SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

DEGENERATE HOPF BIFURCATION AND NERVE IMPULSE. PART II*

ISABEL SALGADO LABOURIAU†

Abstract. The bifurcation from equilibrium of periodic solutions of the Hodgkin and Huxley equations for the nerve impulse is studied. In earlier work singularity theory techniques were used to establish that these equations have a branch of periodic solutions undergoing two Hopf bifurcations, and the equations were conjectured to be equivalent to a member of a one-parameter family of generalized Hopf bifurcation problems. Here the invariants for equivalence to this family and the value of the modal parameter are computed (see [W. W. Farr et al., "Degenerate Hopf bifurcation formulas and Hilbert's 16th problem," *SIAM J. Math. Anal.*, 20 (1989), pp. 13-30]). The value of this parameter determines the type of bifurcation, and in this way it is decided which of the proposed bifurcation diagrams are actually to be found. Thus a topological description of periodic orbits of the Hodgkin and Huxley equations near the equilibrium solution is obtained. In this way, a periodic solution branch is found that does not arise through a classical Hopf bifurcation.

Key words. Hopf bifurcation, Hodgkin-Huxley equations, nerve impulse

AMS(MOS) subject classifications. primary 58F14; secondary 92A09, 58F22

1. Introduction. It is well known [9] that many nerve cells generate trains of impulses as a response to a constant stimulus. In this paper we describe periodic solutions of the clamped Hodgkin and Huxley equations for the nerve impulse (HH), a system of four nonlinear ordinary differential equations, that contain several auxiliary parameters [5]. As in an earlier article [10], here we study the equations HH as a bifurcation problem, regarding the stimulus intensity **I** as a bifurcation parameter. We also describe the dependence of the periodic solutions on some of the other parameters involved in HH, such as the temperature T and the maximum sodium conductance $\bar{g}_{Na}$ that measures the maximum permeability of the nerve cell membrane to Na^+ ions.

Both here and in [10] we have studied the effect of varying $\bar{g}_{Na}$ away from the value, here called *normal*, of 120 m.mho/cm^2 obtained by Hodgkin and Huxley [5] in experiments on the squid giant axon; all other values of $\bar{g}_{Na}$ are called *perturbed* and the HH equations with $\bar{g}_{Na}$ different from normal are referred to as *perturbed* HH. The reader should not be deceived by the word normal; the actual value of $\bar{g}_{Na}$ varies from cell to cell. Hodgkin and Huxley [5] report measurements of $\bar{g}_{Na}$ in the interval [65, 260] m.mho/cm^2 in normal giant squid axons. This variability may be largely due to difficulties in the measurement of $\bar{g}_{Na}$ in experiments, but all the same it makes sense to ask what the equations yield for values different from the average of 120 m.mho/cm^2. The maximum sodium conductance can also be modified by the experimental conditions; low concentrations of local anesthetics or tetrodoxin have the effect of lowering $\bar{g}_{Na}$ and do not seem to affect the conductance of other ions (see [9, Chap. 11]).

In this article, we obtain qualitative amplitude diagrams for the periodic solutions of the perturbed HH (see Figs. 3 and 7) showing the response of a nerve cell to a steadily applied current. This corresponds to a usual experimental procedure ([9] and references therein).

* Received by the editors July 24, 1986; accepted for publication (in revised form) April 6, 1988.

† Grupo de Matemática Aplicada, Faculdade de Ciências, Universidade do Porto, 4000 Porto, Portugal. The work of this author was supported by the Calouste Gulbenkian Foundation and the Instituto Nacional de Investigação Científica of Portugal.

In § 2 we describe the results of [10] on the temperature dependence of the amplitude diagrams and the conjecture presented there concerning the hidden organizing centre for HH; we also give a naive description of the singularity theory concepts used (we refer the reader to [2] for the rigorous construction and for proofs).

The main results are presented in § 3 where the invariants for generalized Hopf bifurcation are evaluated numerically for HH. We determine which of the two possible cases conjectured in [10] takes place and we discuss the behavior near the hidden organizing center, thus obtaining a qualitative description of the amplitude of periodic solutions of HH. We also show that around a critical temperature T_c, the HH equations must have a branch of periodic solutions that does not arise through a classical Hopf bifurcation. These solutions are not easy to find in a numerical integration of the equations—in order to find them we have to know where they are.

Amplitude diagrams for low $\bar{g}_{Na}$ are described in § 4, with a discussion of the possible form of transition from the hidden organizing center to this second family of bifurcation problems.

2. Preliminary results and definitions. In what follows we use the notation and sign conventions of [10]. The perturbed HH equations have a unique temperature-independent, steady-state solution for each value of **I**. This is only true for values of the ionic equilibrium voltages $\bar{V}_{ion}$ close to those determined in [5], as is shown in [6]. After a change of variables, we introduced in [10] a new bifurcation parameter λ, so that for all values of λ the origin of $\mathbb{R}^4$ is a steady-state and $f(\lambda) = \mathbf{I}$ is a monotonically *decreasing* function. In this way we eliminate one error factor in numerical computations since we no longer compute the coordinates of the steady-state as a function of the parameters, and we may think of HH as a family of ordinary differential equations $\dot{y} = \mathrm{HH}(y, \lambda)$ with HH: $\mathbb{R}^5 \to \mathbb{R}^4$ and HH $(0, \lambda) = 0$ where HH also depends on the parameters T and $\bar{g}_{Na}$.

For fixed T and $\bar{g}_{Na}$ and below a critical temperature $T_c(\bar{g}_{Na})$, there are two values $\lambda_1 < \lambda_2$ of the bifurcation parameter where the linearization of HH at the origin has a pair of eigenvalues crossing the imaginary axis transversely. Therefore, by the Hopf bifurcation theorem [7], two distinct periodic solution branches emerge from the equilibrium solution at $(0, \lambda_1)$ and $(0, \lambda_2)$.

For $T > T_c$ no Hopf bifurcation was observed. At $T = T_c$ the two bifurcations coalesce at $\lambda_1 = \lambda_2 = \lambda_c$. We call $(0, \lambda_c)$ a *generalized Hopf bifurcation point*; at this point the linearization of HH has a pair of purely imaginary eigenvalues $\pm i\omega$; all other eigenvalues are real and negative, but some of the other hypotheses in the classical Hopf theorem [7] are violated.

Generalized Hopf bifurcation is studied in [1], [2] (see also [0]), for a system of ordinary differential equations

$$\dot{y} = g(y, \lambda), \quad g(0, \lambda) = 0, \quad y \in \mathbb{R}^n, \quad \lambda \in \mathbb{R} \tag{2.1}$$

such that $D_y g\ (0, 0)$ has simple eigenvalues $\pm i\omega$ and no other eigenvalues of the form $\pm ik\omega$ with $k \in \mathbb{Z}$. The Lyapunov-Schmidt reduction and symmetry considerations are used to represent periodic solutions of (2.1) as the solutions of an equation of the following form:

$$x\mathbf{a}(x^2, \lambda) = 0, \quad \mathbf{a}: \mathbb{R}^2 \to \mathbb{R}, \quad x \geqq 0. \tag{2.2}$$

Intuitively the reduction amounts to rewriting the original vector field locally in "polar coordinates" on a suitable λ-parametrized family of two-dimensional invariant submanifolds of $\mathbb{R}^n$, with periodic orbits corresponding, for each λ, to solutions of $\dot{r} = r\mathbf{a}(r^2, \lambda) = 0$. The steady-state solution is represented by the $x = 0$ solution of (2.2),

and points where periodic orbits bifurcate from it are multiple roots of (2.2) at $x=0$. The set $\{(x, \lambda)\colon x\mathbf{a}(x^2, \lambda)=0\}$ is called a *bifurcation diagram*, and its graph represents a qualitative amplitude diagram for periodic solution branches.

As an example, a vector field (2.1) satisfying the hypotheses of the Hopf theorem [7] at $\lambda=\lambda_c$ after reduction yields a map $\mathbf{a}(x^2, \lambda)$ that for $u=x^2$ satisfies:

$$\mathbf{a}(0, \lambda_c)=0, \quad \mathbf{a}_u=\frac{\partial \mathbf{a}}{\partial u}(0, \lambda_c)\neq 0, \quad \mathbf{a}_\lambda=\frac{\partial \mathbf{a}}{\partial \lambda}(0, \lambda_c)\neq 0. \tag{2.3}$$

Therefore near $(0, \lambda_c)$ there is a unique solution φ to the equation $\mathbf{a}(u, \varphi(u))\equiv 0$ with $\varphi(0)=\lambda_c$, $\varphi'(0)=-\mathbf{a}_u/\mathbf{a}_\lambda$; thus (2.2) has exactly two solution branches through $(0, \lambda_c)$ given by $x=0$ and by $\lambda(x)=\varphi(x^2)$. If $\varphi'(0)<0$, then $\lambda(x)\leqq\lambda_c$ for x near 0 and (2.1) has a periodic solution of nonzero amplitude near $(0, \lambda_c)$ for each $\lambda<\lambda_c$ (Fig. 1(a)). Thus the sign of $\varphi'(0)$ determines the *direction of bifurcation* of periodic solutions of (2.1).

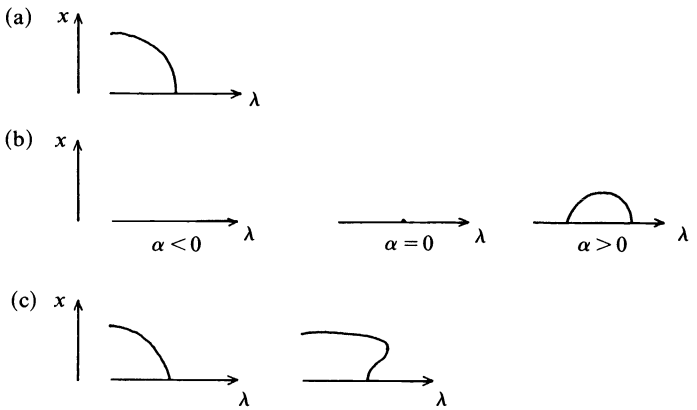


FIG. 1. *Examples of bifurcation diagrams for generalized Hopf bifurcation, ordinates always standing for amplitude* (x): (a) *Classical Hopf bifurcation* (2.3). (b) *Generic perturbation (unfolding) of the codimension* 1 *Hopf bifurcation* (2.7) *in the case* $\mathbf{c}_u>0$, $\mathbf{c}_\lambda=0$, $\mathbf{c}_{\lambda\lambda}>0$. (c) *Generic perturbation (unfolding) of the codimension* 1 *Hopf bifurcation in the case* $\mathbf{h}_u=0$, $\mathbf{h}_\lambda>0$, $\mathbf{h}_{uu}<0$.

The Lyapunov-Schmidt reduction uses the implicit function theorem in a suitable function space and therefore all the results on generalized Hopf bifurcation mentioned here are local both in the variables and in the parameters. In order to avoid repeating phrases like "in some neighborhood of the point $(0, \lambda_c)$," we refer to the *germ of a map* f at a point p—the class of all maps that agree with f in some neighborhood of p. Similarly, we may define the *germ of a set* S at p as the class of all sets that coincide with S in some neighborhood of p.

The germs at $(0, \lambda_c)$ of two maps $\mathbf{a}$, $\bar{\mathbf{a}}$ of the form (2.2), obtained by Lyapunov-Schmidt reduction, are called *contact equivalent* when there are smooth germs $T(x, \lambda)$, $X(x, \lambda)$ and $\Lambda(\lambda)$ transforming one into the other, i.e.,

$$x\mathbf{a}(x^2, \lambda)=T(x^2, \lambda)\circ[x\bar{\mathbf{a}}(X^2(x^2, \lambda)x, \Lambda(\lambda)\lambda)] \tag{2.4}$$

with $T\colon \mathbb{R}^2\to\mathbb{R}$, $T(0, 0)\neq 0$, $X(0, 0)>0$, and $\Lambda(0)>0$. The new parameter Λ does not depend on x, so two systems such as (2.1) reducing to $\mathbf{a}$ and $\bar{\mathbf{a}}$ have the same number of periodic solutions near $y=0$ for corresponding values of λ near λ_c. In other words,

if two equations reduce to contact equivalent maps, the sets of their periodic orbits can be smoothly transformed into each other (see [10]). For instance, any two germs at $(0, 0)$ of maps $\mathbf{a}, \bar{\mathbf{a}}: \mathbb{R}^2 \to \mathbb{R}$ satisfying (2.3) will be contact equivalent, provided their partial derivatives at $(0, 0)$ have the same sign (see [1]).

Another concept used here is that of *universal unfolding* of a germ $\mathbf{a}$, a parametrized family of germs that exhibit all the possible bifurcation diagrams present in a neighborhood of $\mathbf{a}$; the *codimension* of $\mathbf{a}$ is the minimum number of parameters necessary to obtain a universal unfolding. For rigorous singularity theory results and definitions we refer the reader to [1] and [2], where some methods for computing codimensions and unfoldings are given.

Using singularity theory techniques, the germs of problems of the form (2.2) have been classified up to contact equivalence in [1] and [2], where a representative in simple polynomial form is given for each contact class occurring generically in three-parameter families of generalized Hopf bifurcations and for its universal unfolding. Each contact class is characterized in [1] and [2] by necessary and sufficient conditions on the Taylor expansion of $\mathbf{a}(x^2, \lambda)$ around the bifurcation point $(0, \lambda_c)$. Explicit formulae for the calculation of derivatives of $\mathbf{a}$ up to third order from those of the original vector field can be found in [0].

For example, if the germ at $(0, 0)$ of $\mathbf{a}: \mathbb{R}^2 \to \mathbb{R}$ satisfies

$$\begin{aligned} &\mathbf{a}(0,0) = 0 \quad \text{and for } u = x^2, \\ &\mathbf{a}_u = \frac{\partial \mathbf{a}}{\partial u}(0,0) > 0, \quad \mathbf{a}_\lambda = \frac{\partial \mathbf{a}}{\partial \lambda}(0,0) = 0, \quad \mathbf{a}_{\lambda\lambda} = \frac{\partial^2 \mathbf{a}}{\partial \lambda^2}(0,0) > 0, \end{aligned} \tag{2.5}$$

then it can be shown [1], [2] that the germ of $\mathbf{a}(x^2, \lambda)$ is contact equivalent at $(0, 0)$ to

$$\mathbf{c}(x^2, \lambda) = x^2 + \lambda^2. \tag{2.6}$$

Since $\mathbf{c}(x^2, \lambda) = 0$ only when $x = \lambda = 0$, the germ of the solution set of both $x\mathbf{c}\,(x^2, \lambda) = 0$ and of $x\mathbf{a}(x^2, \lambda) = 0$ is the λ-axis. If $\mathbf{a}$ has ben obtained from a system like (2.1), this means there are no periodic solutions near $(0, 0)$.

The contact class of (2.6) is not *structurally stable.* An arbitrarily small perturbation of the form

$$\mathbf{C}_\alpha(x^2, \lambda) = \mathbf{c}(x^2, \lambda) - \alpha = x^2 + \lambda^2 - \alpha, \qquad \alpha \in \mathbb{R} \tag{2.7}$$

is not contact equivalent to $\mathbf{c}$ for $\alpha \neq 0$; the set $\mathbf{C}_\alpha = 0$ is either a circle of radius $\sqrt{\alpha}$ or the empty set. If $\mathbf{C}_\alpha$ is the reduced form of a vector field, then for each $\alpha > 0$ it undergoes two classical Hopf bifurcations at $\lambda = \pm\sqrt{\alpha}$ (see Fig. 1(b)). For $\alpha < 0$ the periodic solution set is still trivial (steady-state only) but $\mathbf{C}_\alpha = 0$ has no solutions, and thus $\mathbf{C}_\alpha$ with $\alpha \neq 0$ is not contact equivalent to $\mathbf{c}$. Moreover, for $\alpha \neq 0$, each germ (2.7) is structurally stable and (2.6) is not. It can be shown [1] that the germ (2.6) has codimension 1 and the family of bifurcation problems $\mathbf{C}_\alpha$ is a universal unfolding for it. The one-parameter family of bifurcation problems (2.7) is structurally stable as a whole, although this is not true of any germ satisfying (2.5).

Let $x\mathbf{h}(x^2, \lambda)$ be the germ of the perturbed HH after reduction. We have computed the first-order derivatives of $\mathbf{h}$ and established the following results, appearing in [10].

(A) At the critical temperature $T_c(\bar{g}_{\mathrm{Na}})$ where the two bifurcation points coalesce, we have $\mathbf{h}_\lambda = \partial/\partial\lambda\, \mathbf{h}(0, \lambda_c) = 0$ for every choice of $\bar{g}_{\mathrm{Na}}$ in the interval [85, 130] (we will omit units from now on). For normal HH if $\mathbf{h}_{\lambda\lambda} > 0$ then $\mathbf{h}$ is equivalent to the generalized Hopf bifurcation germ (2.6). The effect of small temperature variations on the zero

set of $\mathbf{h}$ is the same as that of variations in the value of α over the zero set of (2.7), as in Fig. 1(b). Thus for $T < T_c$, HH has a single stable periodic solution branch undergoing two classical Hopf bifurcations.

(B) For normal HH, the direction of bifurcation changes at one of the Hopf bifurcation points $(0, \lambda_2)$ at a lower temperature, T_1, following a change in the sign of $\partial/\partial u\, \mathbf{h}(0, \lambda_2) = \mathbf{h}_u$. The corresponding periodic solution branch loses stability. If at this temperature $\mathbf{h}_{uu} \neq 0$, then the system is contact equivalent to another generalized Hopf bifurcation germ, also of codimension 1. The study of its universal unfolding, as in the case above, provides the description of the amplitude of periodic solutions (Fig. 1(c) for the case $\mathbf{h}_{uu} < 0$, $\mathbf{h}_\lambda > 0$) as T is varied around T_1. Clearly, the neighborhood of the bifurcation point where the contact equivalence holds does not include the other (nondegenerate) Hopf bifurcation point of HH present at the same temperature.

In both cases above, the study of HH at the degenerate Hopf bifurcation points provides additional information about periodic solutions at nearby parameter values. In case (A) it shows that the two Hopf bifurcations involve the same periodic solution branch, a nontrivial observation in a four-dimensional phase space. In the second case, for $T < T_1$, the classical Hopf bifurcation theorem shows the existence of unstable periodic solutions for $\lambda < \lambda_2$. Thus the presence of a degeneracy not only explains the transition from one type of diagram to another, it also makes the analysis "more global." Here, a point in parameter space where a degenerate Hopf bifurcation occurs will be called an *organizing center* for the equations.

(C) For the perturbed HH the function

$$\bar{g}_{\mathrm{Na}} \to \mathbf{h}_u(\bar{g}_{\mathrm{Na}}) = \mathbf{h}_u(0, \lambda_c) \quad \text{at } T = T_c(\bar{g}_{\mathrm{Na}}) \quad (\text{where } \mathbf{h}_\lambda = 0) \tag{2.8}$$

changes sign twice in the interval [85, 130] and one of its zeros lies within ten percent of the normal value of $\bar{g}_{\mathrm{Na}}$. This point we called a *hidden organizing center* for the equations—it is hidden in the sense that HH had to be perturbed in order to find it.

Even if this value of $\bar{g}_{\mathrm{Na}}$ were never assumed in practice, we would expect the study of the degeneracy to bring together the two descriptions (A) and (B) above, in the same way local information about the two classical Hopf bifurcations in (A) was put together by the study of (2.7). At the hidden organizing center, HH should be equivalent after reduction to a germ containing the two original ones in its universal unfolding, i.e., a family of germs (A_α) such that each first derivative of (A_α) is zero for values of α arbitrarily close to zero. It is easy to see that this is possible only in a structurally stable family if the parameter α is at least two-dimensional.

The simplest (lowest codimension) problems containing in its universal unfolding both germs mentioned in (A) and (B) above are members of a family of codimension 3 problems. The family is characterized by zero first-order derivatives at the organizing center, by the signs of its second-order derivatives, as well as by the value of a modal parameter $\mathbf{b}$. We discussed the geometry of the periodic solution branches for different values of $\mathbf{b}$ when we conjectured in [10] that HH is contact equivalent to this family at the zero of $\mathbf{h}_u(\bar{g}_{\mathrm{Na}})$ closest to the normal value of 120.

In the next section we present the result of the numerical evaluation for HH of the nondegeneracy conditions for equivalence to the one-parameter family discussed above, and of the calculation of the modal parameter $\mathbf{b}$. Our goal is to obtain a description of all nonequivalent bifurcation diagrams that appear in HH at temperatures close to T_c, a subset of those on the universal unfolding of $\mathbf{h}$. As the value of the modal parameter determines the type of bifurcation appearing on the unfolding of $\mathbf{h}$, in this way we decide which of the proposed bifurcation diagrams are actually to be found

in HH. This is discussed in § 3 for the hidden organizing center, and in § 4 for the second zero of $\mathbf{h}_u(\bar{g}_{\mathrm{Na}})$.

3. Diagrams near the hidden organizing center. The hidden organizing center for HH is a point in parameter space where the reduction $\mathbf{h}$ of the perturbed HH satisfies

$$\mathbf{h}(0, \lambda_c) = \mathbf{h}_u(0, \lambda_c) = \mathbf{h}_\lambda(0, \lambda_c) = 0 \quad \text{with } u = x^2. \tag{3.1}$$

We conjectured in [10] that around this point $\mathbf{h}$ is equivalent to a member of the family

$$\mathbf{a}_{\mathbf{b}}(u, \lambda) = \varepsilon u^2 + 2\mathbf{b}\lambda u + \delta\lambda^2 \tag{3.2}$$

studied in [1] and [12]. It is shown in [1] that a germ $\bar{\mathbf{a}}$ is contact equivalent to (3.2) if and only if it satisfies

$$\begin{aligned} &\bar{\mathbf{a}}(0, \lambda_c) = \bar{\mathbf{a}}_u(0, \lambda_c) = \bar{\mathbf{a}}_\lambda(0, \lambda_c) = 0, \\ &\bar{\mathbf{a}}_{uu}(0, \lambda_c) = \bar{\mathbf{a}}_{uu} \neq 0 \neq \bar{\mathbf{a}}_{\lambda\lambda}(0, \lambda_c) = \bar{\mathbf{a}}_{\lambda\lambda}, \quad \text{and} \\ &\mathbf{b} = \bar{\mathbf{a}}_{u\lambda} / |\bar{\mathbf{a}}_{uu}\bar{\mathbf{a}}_{\lambda\lambda}|^{1/2} \neq 0 \\ &\text{with } \delta\mathbf{b}^2 \neq 1 \quad \text{where } \varepsilon = \operatorname{sign}(\bar{\mathbf{a}}_{uu}), \quad \delta = \operatorname{sign}(\bar{\mathbf{a}}_{\lambda\lambda}). \end{aligned} \tag{3.3}$$

The numbers $\varepsilon\delta$ and $\mathbf{B} = \varepsilon\mathbf{b}$ form a complete set of invariants for the family (3.2) under contact equivalence [11], i.e., two members of this family are contact equivalent if and only if $\varepsilon\delta$ and $\mathbf{B}$ are the same for them. The situation differs markedly from the example of § 2, where a similar set of conditions defined a single contact equivalence class. Here each choice of signs in the second-order derivatives of $\bar{\mathbf{a}}$ defines a continuum of nonequivalent bifurcation problems.

Using the methods of [2] it can be shown that (3.2) has universal unfolding given by

$$A(u, \lambda, \alpha, \beta, \mathbf{b}) = \varepsilon u^2 + 2\mathbf{b}\lambda u + \delta\lambda^2 + \operatorname{sign}(\mathbf{b})\beta\lambda + \alpha. \tag{3.4}$$

Therefore, for any sufficiently small perturbation of a germ satisfying (3.3) there are smooth changes of coordinates of the form (2.4) transforming it into one of the germs (3.4) for some choice of α, β, and $\mathbf{b}$. At $\alpha = \beta = 0$ the family (3.4) coincides with (3.2).

The definition of contact equivalence can be weakened so as to obtain a discrete classification of the family (3.2). We define *topological contact equivalence* in the same way as the (smooth) contact equivalence (2.4), with continuity substituted for smoothness (see [2]). Two germs belonging to the same equivalence class (called *modal class*) have homeomorphic bifurcation diagrams.

Let $(3.2)_\pm$ be the family (3.2) with $\varepsilon\delta = +1$ or -1, respectively. Under topological contact equivalence, $(3.2)_+$ splits into four equivalence classes (called *modal classes*), corresponding to values of $\mathbf{B}$ in the intervals $]-\infty, -1[$; $]-1, 0[$; $]0, 1[$, and $]1, +\infty[$. The solutions of $A(u, \lambda, \alpha, \beta, \mathbf{b}) = 0$ are either ellipses or the empty set in the (u, λ)-plane for $0 < |\mathbf{B}| < 1$ (Fig. 2) or hyperbolas for $|\mathbf{B}| > 1$ (Fig. 3). In the unfolding of $(3.2)_-$, only two modal classes occur corresponding to $\mathbf{B} > 0$ and $\mathbf{B} < 0$, and the solution to $A(u, \lambda, \alpha, \beta, \mathbf{b}) = 0$ is always a hyperbola in the (u, λ)-plane (Fig. 7).

By comparing the bifurcation diagrams for (3.4) to the results summarized in § 2 as well as to those in [4] and [13], we conjectured that $\varepsilon\delta = +1$, and $\mathbf{B} < 0$ for HH. Besides checking the conjecture numerically, we obtain here an estimate of $\mathbf{B}$, and thus of the modal class of $\mathbf{h}$.

The derivatives $\mathbf{h}_\lambda$ and $\mathbf{h}_u$ are simultaneously zero for $\bar{g}_{\mathrm{Na}} = \bar{g}_c$ in the interval [109, 110], and around this point we have the values of Table 1.

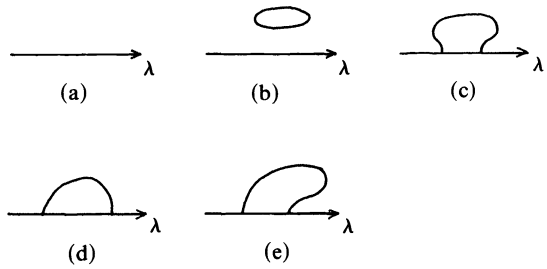


FIG. 2. *Bifurcation (or amplitude) diagrams for stable germs in the universal unfolding* (3.4) *of* $(3.2)_+$ *with* $-1<\mathbf{B}<0$, *for the following parameter values*: (a) $\alpha>\beta^2/4$ *or* $\alpha>\beta^2/4(\mathbf{B}-1)$ *and* $\beta<0$; (b) $\beta>0$ *and* $\beta^2/4(\mathbf{B}-1)<\alpha<\beta^2/4$; (c) $\beta>0$ *and* $0<\alpha<\beta^2/4(\mathbf{B}-1)$; (d) $\beta<0$ *and* $0<\alpha<\beta^2/4(\mathbf{B}-1)$; *and* (e) $\alpha<0$.

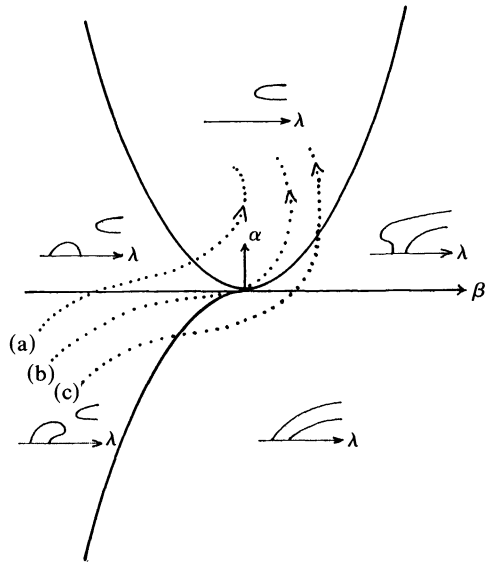


FIG. 3. *The universal unfolding* (3.4) *of* $(3.2)_+$ *with* $B<-1$. *Each bifurcation (or amplitude) diagram is drawn inside the region in* (α,β)*-plane where it occurs*; *the* λ *axis is the steady-state in each case. Dotted lines represent* HH *as temperature-parametrized curves on the unfolding of* $(3.2)_+$, *with temperature increasing as indicated by the arrows*: (a) $\bar{g}_{Na}>\bar{g}_c$; (b) $\bar{g}_{Na}=\bar{g}_c$; and (c) $\bar{g}_{Na}<\bar{g}_c$.

TABLE 1

$\bar{g}_{Na}$ m.mho/cm^2	λ_c mV	$\mathbf{I}(\lambda_c)$ mA/cm^2	$T_c(\lambda_c)$ °C
109	−15.97	74.06	26.278
110	−16.01	74.22	26.546

Numerical estimates of the second-order derivatives of $\mathbf{h}$ at $T_c(\bar{g}_{Na})$ were obtained for several values of $\bar{g}_{Na}$, by using the formulae of [0], a correction to those of § 5 of [1] for the case when the imaginary eigenvalues are $\pm i\omega$, with $\omega\neq 1$. Thus the numerical results of Table 1 differ both from the preliminary calculations presented in [2] and from those in [11], but the qualitative behavior remains the same. For $\bar{g}_{Na}$ in the interval [85, 130] we have found the following:

(D) $\mathbf{h}_{u\lambda}$ is always negative and decreases with $\bar{g}_{Na}$ (Fig. 4).

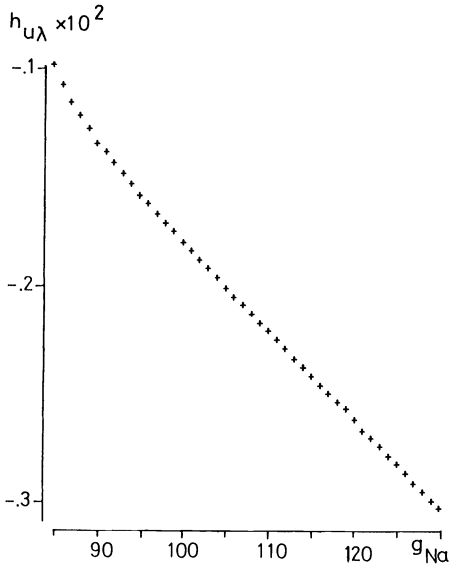


FIG. 4. *The derivative* $\mathbf{h}_{u\lambda}$ *at the point where there is a single generalized Hopf bifurcation, as a function of* $\bar{g}_{Na}$, *for the perturbed Hodgkin and Huxley equations.*

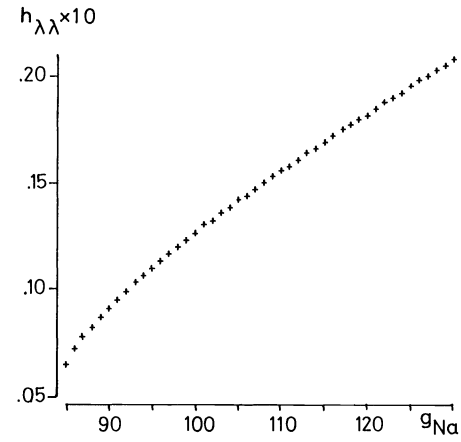


FIG. 5. *The derivative* $\mathbf{h}_{\lambda\lambda}$ *at the point where there is a single generalized Hopf bifurcation, as a function of* $\bar{g}_{Na}$, *for the perturbed Hodgkin and Huxley equations.*

(E) $\mathbf{h}_{\lambda\lambda}$ is always positive and increases with $\bar{g}_{Na}$ (Fig. 5).

(F) $\mathbf{h}_{uu}$ increases with $\bar{g}_{Na}$ and changes sign for $\bar{g}_{Na}$ in the interval]96, 97[(Fig 6).

We have confirmed the following conjecture presented in [10]: at the hidden organizing center, conditions (3.4) are satisfied, and therefore *the perturbed* HH, *in reduced form, are equivalent to a member of the family* (3.2) *with* $\varepsilon = +1 = \delta$.

The modal parameter $\mathbf{b}$ was found to be $\mathbf{b} = -7.0$ at $\bar{g}_{Na} = 109.0$, and $\mathbf{b} = -6.8$ at $\bar{g}_{Na} = 110.0$, where both values are rounded to the number of digits shown, and all digits are believed to be correct. Thus we have also determined that *at the hidden organizing center the reduced* HH *are in the modal class* $\mathbf{B} < -1$ *of the family* $(3.2)_+$.

For $\bar{g}_{Na}$ and T near $\bar{g}_c$ and $T_c(\bar{g}_c)$, respectively, the perturbed HH are equivalent to one of the germs (3.4). In this way we can visualize HH as a surface on $\mathbb{R}^3$ ($(\alpha, \beta, \mathbf{b})$-space) fibered by the T-parametrized curves corresponding to fixed values of $\bar{g}_{Na}$. A zero of $\mathbf{h}_u(\bar{g}_{Na}, T_c(\bar{g}_{Na}))$ will correspond to a curve through the organizing

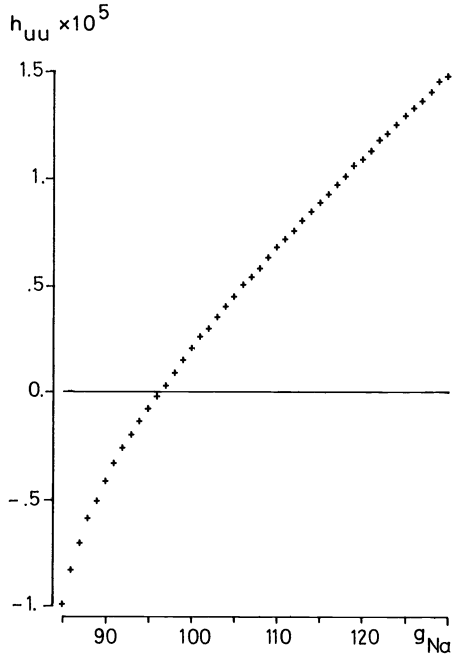


FIG. 6. *The derivative* $\mathbf{h}_{uu}$ *at the point where there is a single generalized Hopf bifurcation, as a function of* $\bar{g}_{Na}$, *for the perturbed Hodgkin and Huxley equations.*

center $(0, 0, \mathbf{b})$, if we recall that $\mathbf{h}_\lambda(\bar{g}_{Na}, T_c(\bar{g}_{Na}))$ is always zero. We have represented in Fig. 3 the perturbed HH as T-parametrized curves on the unfolding of $(3.2)_+$ for $\mathbf{B} < -1$. Each persistent bifurcation diagram is drawn inside the corresponding region in the (α, β)-plane. The **B** direction has been omitted since bifurcation diagrams look the same along it, due to the topological contact equivalence.

The curves representing HH in Fig. 3 were obtained by inspecting the bifurcation diagrams for $(3.2)_+$ and comparing them to data on the direction of bifurcation from [10], [11], and [13] and were confirmed for $\bar{g}_{Na} = 100$.

We find that for $\bar{g}_{Na} = \bar{g}_c$ only two persistent bifurcation diagrams are present. For $T < T_c$ there is a single periodic solution branch undergoing two nondegenerate Hopf bifurcations in the same direction as observed in (B) of § 2; this branch disappears at $T = T_c(\bar{g}_c)$. Both bifurcation diagrams contain a solution branch isolated from the trivial solution $x = 0$.

The two persistent bifurcation diagrams described above occur for normal HH along with a third diagram where the periodic solution branches grow toward each other, corresponding to the situation described in (A) of § 2. In this way our knowledge of the hidden organizing center allows us to put together the partial information of § 2, (A) and (B).

All diagrams for $\bar{g}_{Na} \geqq \bar{g}_c$ contain a branch of solutions not connected to the trivial solution $x = 0$, a characteristic of the modal classes $|\mathbf{B}| > 1$ with $\varepsilon\delta = 1$. Indeed it is natural to find two disconnected branches in some of the bifurcation diagrams since the zeros of $(3.4)_+$ are hyperbolas in the (u, λ)-plane in this case. This is not possible for $|\mathbf{B}| < 1$, where $A(u, \lambda, \alpha, \beta, \mathbf{b}) = 0$ on a bounded set.

The isolated solution branch has never been found in a numerical integration of HH. Since these solutions do not arise through a classical Hopf bifurcation, they could

easily be missed in a numerical tracing of the solutions. It is also possible, however, that at the normal value of $\bar{g}_{Na}$ the isolated branch has either disappeared through some global bifurcation not captured by the Lyapunov-Schmidt reduction, or moved to a region of parameter space without physiological meaning. On the other hand, the presence of this solution branch may explain the damped oscillations observed both numerically [5], [8] and in experiments described in [8], [3].

For $\bar{g}_{Na} < 109$ the following two new types of diagram appear. As T increases the two unconnected branches meet and form two periodic solution branches arising through Hopf bifurcation; a further increase of T yields a change in the direction of bifurcation, before the two Hopf bifurcation points coalesce. This change of direction affects the first Hopf bifurcation point $(0, \lambda_1)$ in contrast to the findings of [10] for normal HH (cf. (B) of § 2). For $\bar{g}_{Na} = 100$ the first-order derivatives of $\mathbf{h}$ at nondegenerate Hopf bifurcation points have been calculated, confirming the change in the direction of bifurcation and thus providing furher evidence for the position of the T-parametrized curve.

Changes in the direction of bifurcation would be observed experimentally as the appearance of hysteresis at the onset of periodic behavior around $\lambda = \lambda_1$. Recall that the parameter λ is a decreasing function of stimulus intensity **I**. When $\bar{g}_{Na} < \bar{g}_c$, a cell under overstimulation would stop its repetitive activity at a value of $\mathbf{I} > f(\lambda_1)$ when **I** increases. If **I** were subsequently decreased, then firing would resume at $\mathbf{I} = f(\lambda_1)$. For $\bar{g}_{Na} \geqq \bar{g}_c$ there is hysteresis at $T < T_c(\bar{g}_{Na})$, around 28°C for normal $\bar{g}_{Na}$. For $\bar{g}_{Na} < \bar{g}_c$, hysteresis is also obtained at high temperature values.

Hysteresis at the offset of repetitive firing for an overstimulated cell has been observed in experiments on squid axons bathed in low Ca^{++} sea water [3]. The reduction in $[Ca^{++}]$ lowers the potential outside the cell membrane and is thus equivalent to the effect of depolarization; a smaller stimulus will be required in order to obtain the same response from the nerve cell [9]. The effective stimulus that can be applied without damaging the cell will be larger and high stimulation effects may become experimentally observable. However, the HH equations have to be modified if they are to describe the new experimental situation [8].

4. Bifurcation diagrams for small $\bar{g}_{Na}$. The derivative $\mathbf{h}_u$, computed for each value of $\bar{g}_{Na}$ at the temperature where the two Hopf bifurcation points coalesce, has a second zero for $\bar{g}_{Na}$ in the interval $]89, 90[$. This singularity is probably of less significance for the nerve impulse than the one of the previous section, since $\bar{g}_{Na}$ is far from the normal value, although it is still physiologically meaningful. At this point, the characteristics are (see Figs. 4–6):

$$\begin{aligned} &\mathbf{h}_u = \mathbf{h}_\lambda = 0, \\ &\mathbf{h}_{uu} < 0, \quad \mathbf{h}_{u\lambda} < 0, \quad \mathbf{h}_{\lambda\lambda} > 0. \end{aligned} \tag{4.1}$$

Again, $\mathbf{h}$ is equivalent to a member of the family (3.2). Around this point the data are as shown in Table 2.

TABLE 2

$\bar{g}_{Na}$ m.mho/cm^2	λ_c mV	$\mathbf{I}(\lambda_c)$ mA/cm^2	$T_c(\lambda_c)$ °C	ε	δ	$\varepsilon\delta$	$\mathbf{B} = \varepsilon\mathbf{b}$
89	−14.61	65.39	17.500	−1	+1	−1	+6.0
90	−14.72	66.22	18.264	−1	+1	−1	+6.8

At this second zero $\bar{g}_p$ of $\mathbf{h}_u$, the perturbed HH are equivalent to a member of the family $(3.2)_-$, in the modal class $\mathbf{B}>0$. The persistent bifurcation diagrams for the germs in the unfolding of $(3.2)_-$ are shown in Fig. 7, where the perturbed HH are represented as a T-parametrized curve on (α, β)-space for values of $\bar{g}_{Na}$ near $\bar{g}_p$.

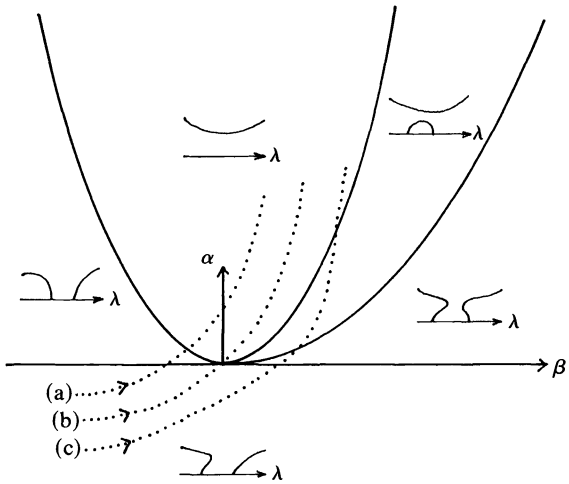


FIG. 7. *The universal unfolding* (3.5) *of* $(3.2)_-$ *with* $\mathbf{B}>0$. *Bifurcation diagrams are drawn inside the regions in* α-β-*space where they occur; the* λ *axis is the steady-state in each diagram. Dotted lines represent* HH *as temperature-parametrized curves on the unfolding of* $(3.2)_-$, *with temperature increasing as indicated by the arrows*: (a) $\bar{g}_{Na}>\bar{g}_p$; (b) $\bar{g}_{Na}=\bar{g}_p$; *and* (c) $\bar{g}_{Na}<\bar{g}_p$.

The diagrams present on the unfolding of $(3.2)_-$ differ markedly from those discussed in § 2. First, most diagrams of Fig. 7 exhibit two Hopf bifurcations that generate disjoint periodic solution branches. Second, for all diagrams in the unfolding of $(3.2)_+$, the set of bifurcation parameter values for which the equation $\mathbf{a}(u, \lambda)=0$ has a nonzero solution is bounded below. This is not true of any diagram in Fig. 7, where in all cases there is a periodic solution branch that can be extended indefinitely in the $\lambda \to -\infty$ direction. These differences may be due to global bifurcations, but it is probable that they are simply an artifact of the local analysis; the methods used here provide a local description of the bifurcation diagrams but give no information as to the size and shape of the neighborhood where they can be applied. It is also clear that as $\bar{g}_{Na}$ increases from 90 to 109, the local behavior of HH is best explained by the analysis of last section.

The situation here is analogous to the results of [10]; initially we knew HH had two classical Hopf bifurcations at $\lambda_1<\lambda_2$ with the same direction of bifurcation and there was no a priori reason for the two periodic solution branches to meet. The proximity of a degeneracy, in the first case a change in the direction of bifurcation as described in (2.3) coinciding with the coalescence of the two bifurcation points given by (2.2), made the analysis "more global." The apparent lack of information about the periodic solution branches away from the bifurcation points was introduced by the local analysis.

The same discussion applies to the two hidden organizing centers presented here. Bifurcation diagrams appear either "capped off" or with an extended branch, depending on whether they are interpreted as part of the unfolding of $(3.2)_+$ or of $(3.2)_-$ (cf. Figs. 3 and 7). Moreover, the unfolding of $(3.2)_+$ contains no germ equivalent to $(3.2)_-$ and vice versa.

Different bifurcation diagrams are obtained for (2.1) when it is studied on its own or as part of the unfolding of a more degenerate (i.e., higher codimension) germ. The presence of a degeneracy close by is extra information about the germ, and this is reflected in the diagrams we obtain for it. The only codimension 3 germs whose universal unfolding contains germs equivalent to (3.2) with $\mathbf{B} \to \pm\infty$ as they approach the origin of the unfolding space are defined by the following conditions (see [2]):

$$\begin{aligned} &\mathbf{a}_{uu} = \mathbf{a}_u = \mathbf{a}_\lambda = 0, \\ &\mathbf{a}_{u\lambda} < 0, \quad \mathbf{a}_{\lambda\lambda} > 0, \quad \mathbf{a}_{uuu} \neq 0, \end{aligned} \tag{4.2}$$

where each sign of $\mathbf{a}_{uuu}$ corresponds to a contact equivalence class.

It may be possible to perturb HH so as to satisfy (4.2), but the perturbation may introduce some further degeneracy into the problem. This is particularly crucial in the present case, since the zero of $\mathbf{a}_{uu}$ is very close to the minimum of the function $\bar{g}_{\mathrm{Na}} \to \mathbf{h}_u(\bar{g}_{\mathrm{Na}})$ defined in (2.8).

The comments made above for the second zero of $\mathbf{h}_u$ apply equally well to this hypothetical organizing center. Given the large number of parameters involved, it is natural to expect HH to have a high codimension. Calculations become more complicated as the codimension increases (see [0]), and therefore it only makes sense to look for a second organizing center if evidence of low $\bar{g}_{\mathrm{Na}}$ persistence of isolated periodic branches is found.

Acknowledgment. We thank W. Farr for the important correction mentioned in § 3.

REFERENCES

[0] W. FARR, C. LI, I. S. LABOURIAU, AND W. F. LANGFORD, *Degenerate Hopf bifurcation formulas and Hilbert's 16th problem*, SIAM J. Math. Anal., this issue (1989), pp. 13-30.

[1] M. GOLUBITSKY AND W. F. LANGFORD, *Classification and unfoldings of Hopf bifurcations*, J. Differential Equations, 41 (1981), pp. 375-415.

[2] M. GOLUBITSKY AND D. SCHAEFFER, *Singularities and Groups in Bifurcation Theory* I, Springer-Verlag, New York, 1985.

[3] R. GUTTMAN, S. LEWIS, AND J. RINZEL, *Control of repetitive firing in squid axon membrane as a model for a neuroneoscillator*, J. Physiol., 305 (1980), pp. 377-395.

[4] B. HASSARD, *Bifurcation of periodic solutions of the Hodgkin-Huxley model for the squid giant axon*, J. Theoret. Biol., 71 (1978), pp. 401-420.

[5] A. L. HODGKIN AND A. F. HUXLEY, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 500-544.

[6] A. HOLDEN, P. G. HAYDON, AND W. WINLOW, *Multiple equilibria and exotic behaviour in excitable membranes*, Biol. Cybernet., 46 (1983), pp. 167-172.

[7] E. HOPF, *Abzweizung einer periodischen Lösung eines Differentialsystems*, Ber. Math. Kl. Säcks. Akad. Wiss. Leipzig, 94 (1942), New York, pp. 3-22.

[8] A. F. HUXLEY, *Ion movements during nerve activity*, Ann. New York Acad. Sci., 81 (1959), pp. 221-246.

[9] J. J. B. JACK, D. NOBLE, AND R. W. TSIEN, *Electric Current Flow in Excitable Cells*, Clarendon Press, Oxford, 1982.

[10] I. S. LABOURIAU, *Degenerate Hopf bifurcation and nerve impulse*, SIAM J. Math. Anal., 16 (1985), pp. 1121-1133.

[11] ———, *Applications of singularity theory to neurobiology*, Ph.D. thesis, University of Warwick, Coventry, United Kingdom, 1983.

[12] ———, *Note on the unfolding of degenerate Hopf bifurcation germs*, J. Differential Equations, 57 (1985), pp. 436-439.

[13] J. RINZEL AND R. N. MILLER, *Numerical solutions of the Hodgkin-Huxley equations*, Math. Biosci., 49 (1980), pp. 27-59.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

DEGENERATE HOPF BIFURCATION FORMULAS AND HILBERT'S 16th PROBLEM*

W. W. FARR†, CHENGZHI LI‡, I. S. LABOURIAU§, AND W. F. LANGFORD¶

Abstract. This paper presents explicit formulas for the solution of degenerate Hopf bifurcation problems for general systems of differential equations of dimension $n \geqq 2$, with smooth vector fields. The main new result is the general solution of the problem for a weak focus of order 3. For bifurcation problems with a distinguished parameter, we present the formulas for the defining conditions of all cases with codimension $\leqq 3$. The formulas have been applied to Hilbert's 16th problem, yielding a new proof of Bautin's theorem, and correcting an error in Bautin's formula for the third focal value. The approach used is the Lyapunov-Schmidt method. A review of five other approaches is given, along with literature references and comparisons to the present work.

Key words. Hopf bifurcation, Lyapunov-Schmidt reduction, Hilbert's 16th problem

AMS(MOS) subject classifications. 58F14, 34C25

1. Introduction.

The Hopf Bifurcation Theorem has become the standard tool in applied mathematics for the study of the birth (or death) of a periodic solution of a differential equation at an equilibrium point. This is fitting, since the classical Hopf theorem states generic conditions on the differential equation for this bifurcation to occur, and also provides an explicit formula (or algorithm) for calculating the periodic solution and its stability. However, there is growing interest in degenerate cases, to which the classical Hopf formula does not apply. This paper is concerned with degeneracies in which multiple periodic solutions may coexist, as well as degeneracies in the dependence on a parameter. Explicit formulas are given for the defining conditions of all degeneracies having codimension $\leqq 3$, as defined in Golubitsky and Langford [14].

The main new result of this paper is the general solution of the focal values of a weak focus of order 3, for a system of dimension $n \geqq 2$, and an arbitrary smooth vector field. A weak focus of order 3 implies the existence of up to three coexisting limit cycles, under small perturbations. Previously, only the focus of order 2 had been solved in this generality, and work on the weak focus of order 3 had been restricted to special cases, such as dimension $n = 2$ and quadratic nonlinearities.

The formulas presented here are applicable to investigations of oscillations in chemical reactions (Farr and Aris [12]), biological systems (Labouriau [21], [22]), and many other fields. Admittedly, the formulas are long and complex. The authors feel that the effort involved in deriving and checking them is justified by the generality of the results. The formulas are explicit in terms of the Taylor coefficients of the original vector field. No preliminary transformations or reductions of the differential equations are required; for example, the equilibrium is not required to be "trivial" (i.e., identically zero), and the basic frequency need not be scaled to 1. The formulas can be programmed

* Received by the editors April 1, 1988; accepted for publication April 6, 1988.

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

‡ Department of Mathematics, Beijing University, Beijing, People's Republic of China. The work of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

§ Grupo de Matemática Aplicada, Universidade do Porto, 4000 Porto, Portugal. The work of this author was supported by the Instituto Nacional de Investigação Científica-Portugal.

¶ Department of Mathematics, University of Guelph, Guelph, Ontario, Canada N1G 2W1. The work of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

in a symbolic computation language, such as MACSYMA or Maple, to compute the bifurcation coefficients for specific examples conveniently and with minimal risk of error.

As a nontrivial application of these formulas, we consider the second part of Hilbert's 16th problem, concerning the number and position of limit cycles of planar (two-dimensional) differential equations with polynomial nonlinearities. This problem is still unsolved, even in the quadratic case (see the surveys of Coppel [9], [10], and references therein). Locally, the quadratic case was studied by Bautin in 1952 (see Bautin [4]), who proved, after a long calculation by the succession function method, that in a neighborhood of an equilibrium point of a planar quadratic system, there can be 0, 1, 2, or 3 limit cycles and no more; otherwise the equilibrium point is a center. This paper presents the first new derivation of Bautin's results and corrects an error in one of Bautin's coefficients.

The approach used in this paper is the Lyapunov–Schmidt method, also known as the Fredholm Alternative Method, or Method of Alternative Problems. It is essentially the same method used by Hopf in his classic paper [18]. Recently, Golubitsky and Langford [14] combined singularity theory with the Lyapunov–Schmidt method to classify degenerate Hopf bifurcations and their unfoldings. They gave formulas for the bifurcation coefficients for some degenerate cases. This paper extends those results, giving formulas that are more readily applicable, and including the case of a weak focus of order three.

In the existing literature on degenerate Hopf bifurcation, we can identify six different methods of solution. These are: the method of Poincaré–Birkhoff normal forms; the method of Lyapunov constants; the method of the succession function; the method of averaging; the method of intrinsic harmonic balancing; and the Lyapunov–Schmidt method, which is the one used here. Unfortunately the literature on the six methods is nearly disjoint; there have been no comparative studies to guide the user to a choice among these methods. Therefore, in this paper we briefly describe all of the methods and relate them to the approach used here.

The plan of the paper is as follows. Section 2 briefly reviews the classical Hopf theorem, and defines and compares the six different methods listed above for degenerate Hopf bifurcation problems. Section 3 outlines the calculation and presents the new formulas for the degenerate Hopf bifurcation coefficients. Section 4 applies these formulas to a new proof of Bautin's theorem for Hilbert's 16th problem.

2. Degenerate Hopf bifurcations: alternative approaches. This section reviews six different methods that have been used in studies of degenerate Hopf bifurcation problems. Salient features of each of the methods are described; however, this review is not intended to be an exhaustive comparative analysis. References to more complete presentations of each of the methods are given. Most authors in the field have focused on only one of these methods and have not related their work to the other approaches. We hope this review may help reduce this fragmentation of the field and facilitate comparison of the results presented here with previous work, for example, on Hilbert's 16th problem.

First it is necessary to establish more precisely some terminology and notation, and to briefly review the classical Hopf theorem. Let us consider a parametrized family of differential equations

$$u' = f(u, \mu) \tag{2.1}$$

where $f: \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$, and f is a smooth (C^∞) function of the state variables u and parameters μ, and $n \geqq 2$, $p \geqq 1$. For some applications it is important to preserve a

physical control parameter λ, in which case we define

$$\mu = (\lambda, \alpha). \tag{2.2}$$

Then λ represents a distinguished bifurcation parameter, as in [14], and α represents additional "unfolding" or free parameters; not all authors make this distinction. We assume the existence of an equilibrium point, translated to the origin for convenience,

$$f(0, 0) = 0 \tag{2.3}$$

and at this equilibrium, the derivative with respect to u,

$$A \equiv (D_u f)|_{(0,0)} \tag{2.4}$$

has a simple pair of imaginary eigenvalues $\pm i\omega_0$, and no others with zero real part. Then A is nonsingular, and the Implicit Function Theorem implies that the equilibrium (2.3) has a smooth extension $u(\mu)$ for μ near zero. Usually it is assumed that this branch of equilibria has been translated to the trivial solution $u \equiv 0$; however, to facilitate applications, we do not make this assumption here. Now the derivative A has a smooth extension along the branch of equilibria,

$$A(\mu) \equiv (D_u f)|_{(u(\mu),\mu)}, \tag{2.5}$$

and the simplicity of the imaginary eigenvalues implies that they too have smooth extensions along the branch, for μ near 0,

$$\sigma(\mu) \pm i\omega(\mu), \quad \sigma(0) = 0, \quad \omega(0) = \omega_0. \tag{2.6}$$

The classical Hopf theorem assumes the Hopf transversality or crossing condition, with respect to a distinguished parameter λ [18]

$$\sigma_\lambda(0) \neq 0. \tag{2.7}$$

(The consequences of the failure of this hypothesis have been explored in [14].) Then the main conclusion of the Hopf theorem is that there is a unique (up to phase) smooth branch of periodic solutions, in any small neighborhood of $u = 0$, which can be parametrized by an amplitude of the periodic solution, here denoted r, and there is a smooth relationship between the parameters μ and the amplitude r, of the form

$$r[c_0 + c_2 r^2 + c_4 r^4 + \cdots] = 0, \tag{2.8}$$

where $\cdots$ denotes higher-order terms in r. The coefficients c_0, c_2, etc. are functions of μ. Hopf showed that (2.7) is equivalent to

$$(c_0)_\lambda(0) \neq 0, \tag{2.9}$$

so that, by the Implicit Function Theorem, (2.8) can be solved uniquely for λ as a function of $r > 0$ near the origin. Hopf [18] also presented a formula for $c_2(0)$, in terms of the Taylor coefficients of f. The nondegenerate (or classical) case of Hopf bifurcation is that for which (2.7) holds and in addition

$$c_2(0) \neq 0. \tag{2.10}$$

Then it is clear that (2.8) near (0, 0) defines a curve, asymptotically parabolic in shape, with a unique $r > 0$ for each λ of one sign, and no r for λ of the opposite sign. This curve is the classical Hopf bifurcation diagram. Additional parameters α do not qualitatively change this picture when (2.7) and (2.10) hold, but become important if either condition fails.

The computation of the coefficients $c_2, c_4, \cdots$ is one principal theme of this paper. In fact every one of the methods reviewed here can be described as a means of arriving at an equation of the form (2.8), and then computing the coefficients c_{2j}.

Degenerate Hopf bifurcation occurs when either of conditions (2.7) or (2.10) fails. In this paper we are concerned mainly with (2.10). Then the problem is to calculate the first nonvanishing c_{2k} in (2.8) (only even powers of r appear in any of the methods). Adopting the terminology used in much of the literature cited below, we will say that the differential equation has a *weak focus of order k*, if

$$c_0(0) = \cdots = c_{2(k-1)} = 0, \qquad c_{2k} \neq 0, \tag{2.11}$$

and the constant $c_{2j}(0)$ is the jth *focal value*. When (2.11) holds we obtain multiple periodic solutions from the following standard result.

THEOREM 2.1. *Suppose that the differential equation* (2.1) *has a weak focus of order k. Then by generic perturbations involving k parameters in the differential equation, it is possible to obtain j limit cycles, in a neighborhood of* $u = 0$, *for each j satisfying* $1 \leqq j \leqq k$, *but not for any* $j > k$.

A proof of Theorem 2.1 using singularity theory is given in [14], where a more explicit prescription of the "generic perturbations," as well as the effects of a distinguished parameter, can be found. Corresponding proofs for the other approaches to degenerate Hopf bifurcation can be found in the references given below.

Several of the methods to be described here are applicable only in the two-dimensional case. In principle, under the above hypotheses, the Center Manifold Theorem can be used to reduce a system of dimension $n > 2$ to a planar system. However, in practice this reduction is rarely easy to carry out explicitly, so such methods are severely limited. Let us proceed to the review of the methods.

Method of Poincaré-Birkhoff normal forms. This is one of the best-known methods. Excellent references are the books of Arnold [2] and Guckenheimer and Holmes [16]. After a reduction to two dimensions, a sequence of near-identity nonlinear transformations brings the differential equation to a normal form. Written in amplitude-phase coordinates (r, θ), the differential equation for r has right-hand side of the form (2.8), up to a finite order, and the equation for θ has a similar form except it is even in r instead of odd. Higher-order terms do not have this symmetry, but it can be shown that the nonsymmetric higher-order terms are not important, and Theorem 2.1 holds. It should be noted that the reduction to two dimensions and the transformation to normal form can be combined into one calculation for greater efficiency (see Bibikov [5], Coullet and Spiegel [11]). This method has been applied to Hilbert's 16th problem by Rousseau [28], who found that the calculation of the third focal value $c_6(0)$ for a particular example, using MACSYMA, strained the memory capacity of a modern minicomputer.

Method of Lyapunov constants. A good reference for this method is Göbber and Willamowski [13]. We begin with a two-dimensional system, and instead of transforming the system, we construct a positive definite Lyapunov functional $V(u, \mu)$ (as in Lyapunov's method for asymptotic stability) for which the derivative along trajectories is

$$V = \nabla V \cdot f = \sum v_j r^{2j}. \tag{2.12}$$

The coefficients v_j are called the Lyapunov constants and are functions of the parameters. There is an algorithm for constructing V for which the leading v_j's are zero. The level curves of finite truncations of the series (2.12) define Poincaré-Bendixson domains, from which the existence of limit cycles is obtained. Recently, Bonin and Legault [6]

have shown that (2.12) is equivalent to (2.8), and have verified that Theorem 2.1 holds for this approach. They estimate that this approach is more efficient than the method of Poincaré–Birkhoff normal forms, mainly because it involves computing one Taylor series instead of two. The method was applied to the quadratic Liénard equation (a subcase of Hilbert's 16th problem) by Kohda, Imamura, and Oono [20].

Method of the succession function. A thorough exposition of this method can be found in the book of Andronov, Leontovich, Gordon, and Maier [1]. Again we assume a two-dimensional system, with a weak focus at the origin. We select a ray from the origin (typically the x-axis) and choose initial points on this ray. Sufficiently near the weak focus, the Poincaré map, which follows a solution from the ray back to the ray, is well defined. This map is called the succession function. Locally it can be expanded in a Taylor series in the coordinate along the ray, and this Taylor series can be put in the form (2.8). Zeros of the succession function correspond to periodic solutions, and again Theorem 2.1 holds for the succession function. This is the method used by Bautin [4] on Hilbert's 16th problem, and he proved Theorem 2.1 for that case. A formula for the first focal value, derived by this method, is given in Andronov et al. [1].

Method of averaging. There are many good references for the method of averaging as applied to bifurcation problems (see Chow and Hale [8], Guckenheimer and Holmes [16], Sanders and Verhulst [29]). In a two-dimensional system near a weak focus, the phase angle θ is a strictly monotone function of time t. Therefore it is possible to transform the independent variable from t to θ, and reduce the dimension of the system by one, but the new system is 2π-periodic in θ, instead of autonomous. Integral averaging now leads to a vector field of the form (2.8), and Theorem 2.1 applies.

Method of intrinsic harmonic balancing. For this method, refer to Huseyin and Yu [19] and the references therein. Harmonic balancing involves formally expanding a trial solution in a Fourier series and matching coefficients. Certain inconsistencies that arise in the naive approach are overcome in the method of intrinsic harmonic balancing. It has only recently been applied to degenerate Hopf bifurcation problems.

Method of Lyapunov–Schmidt. We summarize the Lyapunov–Schmidt method, as used in [14]. For a thorough exposition, see Golubitsky and Schaeffer [15] and Vanderbauwhede [31]. The first step is to rescale the time in (2.1), to make the period constant and equal to 2π, by

$$s = \omega_0(1+\tau)t, \tag{2.13}$$

where the new parameter τ is the correction to the period, and is to be determined. Then (2.1) is rewritten

$$N(u, \lambda, \alpha, \tau) \equiv -\omega_0(1+\tau)\frac{du}{ds} + f(u, \lambda, \alpha) = 0, \tag{2.14}$$

and we seek solutions to (2.14) in the space $C^1_{2\pi}$ of continuously differentiable, 2π-periodic vector-valued functions.

The reader is warned that N in (2.14) is defined with the opposite sign to the corresponding N in [14]. As a result, many of the formulas in this paper have differences in sign from the formulas in the earlier paper. The motivation for this change is that it leads to eigenvalue and focal values with the same signs as in the traditional approaches described above, and so comparisons are made easier. We obtain the usual correspondence that negative values imply stability. This is a consequence of the fact that in (2.14), N and f have the same sign, so that eigenvalues and focal values computed from N have the same sign as if they were computed directly from the vector field f.

The linearization of (2.14) is

$$Lu \equiv \left[-\omega_0 \frac{d}{ds} + A\right] u = 0, \tag{2.15}$$

with A defined by (2.4). The kernel of L is spanned by

$$\phi_1(s) = \mathrm{Re}\,(c\,e^{is}), \qquad \phi_2(s) = \mathrm{Im}\,(c\,e^{is}) \tag{2.16}$$

where c is an eigenvector of A satisfying

$$Ac = i\omega_0 c, \qquad c^* c = 2. \tag{2.17}$$

Here * denotes complex conjugate transpose. Similarly, we define adjoint eigenfunctions ψ_1 and ψ_2 as the real and imaginary parts of de^{is}, where d^* is the left eigenvector satisfying

$$d^* A = i\omega_0 d^*, \qquad d^* c = 2. \tag{2.18}$$

The vectors c and d are needed in the formulas in the next section.

The Lyapunov-Schmidt method proceeds by projecting (2.14) onto Ker (L) and a complement, solving on the complement where L is invertible, and substituting that solution into the equation in the kernel. The result is a two-dimensional map, called the bifurcation equation,

$$g : \mathbb{R}^2 \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^2, \qquad g(x, y; \mu, \tau) = 0. \tag{2.19}$$

From the phase-shift symmetry of the periodic functions g is S^1 equivariant. We may arbitrarily fix the phase and define an amplitude, for example, $y = 0$ and $x \geqq 0$. Then the bifurcation equation has the form

$$g_1 = p(x^2, \mu, \tau)x = 0, \qquad g_2 = q(x^2, \mu, \tau)x = 0, \tag{2.20}$$

where $p(0) = q(0) = 0$, and $q_\tau(0) = \omega_0 \neq 0$. Periodic solutions correspond to nontrivial solutions of $p = q = 0$. By the Implicit Function Theorem, the q-equation can be solved for τ, which we then substitute into the p-equation, which gives the final equation

$$a(x^2, \mu)x = p(x^2, \mu, \tau(x^2, \mu))x = 0. \tag{2.21}$$

Now (2.21) has the form (2.8), and it has been shown that Theorem 2.1 holds once again. It remains to calculate the Taylor coefficients of a, which include the $c_{2j}(0)$ in (2.8). This is accomplished by repeated implicit differentiation of (2.20)-(2.21), and the solution of linear differential equations in the complementary space. The results are presented in the next section.

Recall that $\mu = (\lambda, \alpha)$, where λ is a distinguished parameter. The possible degeneracies involving the distinguished parameter λ have not been discussed here, but are explored thoroughly in Golubitsky and Langford [14].

3. The bifurcation coefficients. In this section we present the explicit formulas for the Hopf bifurcation coefficients, for the cases of weak foci up to order 3 (as defined in § 2). The general solution for a weak focus of order 2 has been obtained by Golubitsky and Langford [14], and earlier by Hassard and Wan [17]. The weak focus of order 3 seems not to have been solved before, except in special cases; for example, Bautin [4] treated the case of a two-dimensional system with only quadratic nonlinearities.

Furthermore, the formulas presented here remove two simplifying assumptions that were made in [14]. From the theoretical point of view of that paper, there is no loss of generality in assuming that the imaginary eigenvalues are scaled to $\pm i$ and that even as μ varies the equilibrium solution remains fixed at the origin (i.e., is a trivial solution). In applications, however, these simplifying assumptions are rarely satisfied. The necessary extensions to the general case are presented here. Note also that our formulas include those necessary to determine degeneracies involving the distinguished parameter as defined in [14], as well as the focal values discussed in § 2. These formulas allow the calculation of the defining conditions of all possible degeneracies of codimension up to 3, as defined in [14].

To avoid repetition, it is assumed that the reader is familiar with the notation and formulas in § 5 of [14]. However, the reader is warned that, since the nonlinear operator N is defined here with a sign opposite to that in the earlier paper (for reasons explained in § 2), some of the formulas have signs reversed.

First we provide the formulas that remove the assumption of a trivial solution. Let $f(u, \lambda, \alpha)$ be the right-hand side of our set of ordinary differential equations, and assume that it does not have a trivial solution. (In this section we will suppress the dependence of f on α, the vector of free parameters, for convenience.) We define a function G that has a trivial solution locally by

$$G(v, \lambda) = f(v + \hat{u}(\lambda), \lambda) \tag{3.1}$$

where $\hat{u}(\lambda)$ satisfies

$$f(\hat{u}(\lambda), \lambda) = 0, \tag{3.2}$$

for λ in some neighborhood of the Hopf bifurcation point. That is, $\hat{u}(\lambda)$ is the steady-state solution written as a function of the distinguished parameter λ. The formulas we will present below for the derivatives of the functions p and q, as defined in (2.20), will be in terms of G; we present immediately below the formulas relating derivatives of G to derivatives of the original vector field f. Our purpose in splitting up the calculations in this fashion is twofold. First, the subsequent formulas are simplified to a certain extent, and second, some applications will have trivial solutions and the most general formulas would entail quite a few needless calculations. Note first of all that $D_v^k G = D_v^k f$ for all k, where by $D_v^k G$ we mean the kth derivative of G with respect to v thought of as a symmetric k-linear form. Thus if all we seek is the focal values of § 2, it is irrelevant whether or not f has a trivial solution. It is only when derivatives with respect to the distinguished bifurcation parameter occur that complications arise. The extension of our notation to mixed v and λ derivatives is trivial, since we are considering λ to be a single real parameter. The formulas we will need are given below (in our calculations below all derivatives are to be evaluated at the Hopf point, so we will supply as arguments only the k vectors on which the form is to act):

$$D_v D_\lambda G(x) = D_v^2 f(v_1, x) + D_v D_\lambda f(x),$$

$$D_v^2 D_\lambda G(x, y) = D_v^3 f(v_1, x, y) + D_v^2 D_\lambda f(x, y),$$

$$D_v D_\lambda^2 G(x) = D_v^3 f(v_1, v_1, x) + D_v^2 f(v_2, x) + 2D_v^2 D_\lambda f(v_1, x) + D_v D_\lambda^2 f(x),$$

$$D_v^3 D_\lambda G(x, y, z) = D_v^4 f(v_1, x, y, z) + D_v^3 D_\lambda f(x, y, z),$$

$$D_v^2 D_\lambda^2 f(x, y) = D_v^4 f(v_1, v_1, x, y) + D_v^3 f(v_2, x, y) + 2D_v^3 D_\lambda f(v_1, x, y)$$

$$
\begin{aligned}
&\qquad + D_v^2 D_\lambda^2 f(x, y),\\
D_v D_\lambda^3 G(x) &= D_v^4 f(v_1, v_1, v_1, x) + 3D_v^3 f(v_1, v_2, x) + D_v f(v_3, x)\\
&\quad + 3D_v^3 D_\lambda f(v_1, v_1, x) + 3D_v^2 D_\lambda f(v_2, x) + 3D_v^2 D_\lambda^2 f(v_1, x)\\
&\quad + D_v D_\lambda^3 f(x),\\
D_v^4 D_\lambda G(x, y, z, w) &= D_v^5 f(v_1, x, y, z, w) + D_v^4 D_\lambda f(x, y, z, w),\\
D_v^3 D_\lambda^2 f(x, y, z) &= D_v^5 f(v_1, v_1, x, y, z) + D_v^4 f(v_2, x, y, z) + 2D_v^4 D_\lambda f(v_1, x, y, z)\\
&\quad + D_v^3 D_\lambda^2 f(x, y, z),\\
D_v^5 D_\lambda G(x, y, z, w, t) &= D_v^6 f(v_1, x, y, z, w, t) + D_v^5 D_\lambda f(x, y, z, w, t),
\end{aligned}
\tag{3.3}
$$

where x, y, z, w, and t are vectors in $\mathbb{R}^n$, and the quantities v_i are derivatives of $\hat{u}(\lambda)$ defined by

$$
\begin{aligned}
v_1 &= -(D_v f)^{-1}(D_\lambda f),\\
v_2 &= -(D_v f)^{-1}(D_v^2 f(v_1, v_1) + 2D_v D_\lambda f(v_1) + D_\lambda^2 f),\\
v_3 &= -(D_v f)^{-1}(D_v^3 f(v_1, v_1, v_1) + 3D_v^2 f(v_1, v_2)\\
&\quad + 3D_v^2 D_\lambda f(v_1, v_1) + 3D_v D_\lambda f(v_2) + 3D_v D_\lambda^2 f(v_1) + D_\lambda^3 f).
\end{aligned}
\tag{3.4}
$$

Next we present the formulas for the derivatives of $a(z, \lambda)$ in terms of the derivatives of p and q (we let $z \equiv x^2$, see (2.20)-(2.21)). The notation is that of [14], that is, we write

$$a(z, \lambda) = \sum a_{jk} z^k \lambda^j. \tag{3.5}$$

Note that the derivatives of $a(z, \lambda)$ with respect to z, at $(z, \lambda) = (0, 0)$, are equivalent to the focal values defined in § 2.

The p_{ijk} notation is a shorthand defined like that for the a_{ij} coefficients, which we give below for completeness, although we assume readers of this paper are familiar with either the paper of Golubitsky and Langford [14] or the treatment given in Golubitsky and Schaeffer [15]. The notation depends on the two functions g_1 and g_2 defined above in § 2 and the series expression

$$p(z, \lambda, \tau) = \sum p_{ijk} z^i \lambda^j \tau^k \tag{3.6}$$

where $z = x^2$ as before. Hence p_{ijk} is given by

$$p_{ijk} = \frac{1}{(2i+1)!\, j!\, k!} \frac{\partial^{2i+1+j+k} g_1(0, 0, 0)}{\partial x^{2i+1}\, \partial \lambda^j\, \partial \tau^k} \tag{3.7}$$

with analogous expressions for q_{ijk} involving g_2. The formulas for the a_{ij} coefficients are given below. Their derivation involves only elementary calculus, but they are given here because there are sign changes from [14] due to the new definition of N and also because they are for the case $\omega_0 \neq 1$:

$$
\begin{aligned}
&a_{00} = 0, \quad a_{10} = p_{100}, \quad a_{01} = p_{010},\\
&a_{20} = p_{200} - p_{101}\, q_{100}/\omega_0, \quad a_{11} = p_{110} - p_{101}\, q_{010}/\omega_0, \quad a_{02} = p_{020},\\
&a_{30} = p_{300} - p_{201}\, q_{100}/\omega_0 + p_{102}\, q_{100}^2/\omega_0^2 - p_{101}(q_{200}/\omega_0 - q_{101}\, q_{100}/\omega_0^2)\\
&a_{21} = p_{210} - p_{111}\, q_{100}/\omega_0 - p_{201}\, q_{010}/\omega_0 + 2p_{102}\, q_{100}\, q_{010}/\omega_0^2\\
&\qquad - p_{101}(q_{110}/\omega_0 - q_{101}\, q_{010}/\omega_0^2)\\
&a_{12} = p_{120} - p_{111}\, q_{010}/\omega_0 + p_{102}\, q_{010}^2/\omega_0^2 - p_{101}\, q_{020}/\omega_0 - p_{021}\, q_{100}/\omega_0\\
&a_{03} = p_{030} - p_{021}\, q_{010}/\omega_0.
\end{aligned}
\tag{3.8}
$$

Finally we present the formulas for the derivatives of p and q. Deriving these formulas is straightforward though tedious, and the labor involved especially in the p_{m00} coefficients increases rapidly with m. The formulas assume that G has a trivial solution, so for where this is not true and where we are using a distinguished bifurcation parameter, they should be modified according to the prescription above.

Computation of the coefficients proceeds in two steps. In the first step, linear algebra problems must be solved to obtain complex-valued vectors that actually are coefficients in a Fourier series for the function $w(x, \lambda, \tau)$, and in the second, these vectors are used to evaluate the p and q coefficients of the bifurcation equations. As shown in Golubitsky and Langford [14], certain coefficients are identically zero or have special values. These are

$$\begin{aligned} &p_{00j}=0, \quad j=0,1,2,\cdots, \qquad q_{00j}=0, \quad j=2,3,4,\cdots, \\ &q_{000}=0, \qquad q_{001}=\omega_0, \\ &p_{01j}=q_{01j}=0, \quad j=1,2,3,\cdots. \end{aligned} \tag{3.9}$$

The formulas for the first-order derivatives of $a(z, \lambda)$ at the origin have very simple forms. The set of vectors for the p_{100} coefficient, which determines stability in the case of nondegenerate Hopf bifurcation, is found by solving

$$Aa_0=-\tfrac{1}{2}D_v^2G(c,\bar{c}), \qquad (A-2i\omega_0 I)a_2=-\tfrac{1}{4}D_v^2G(c,c) \tag{3.9}$$

where A is the Jacobian matrix of F at the Hopf point and c is the right eigenvector of A corresponding to the eigenvalue $i\omega_0$, as in § 2. In computations it is useful to form the intermediate PQ100 and PQ010 quantities, given by

$$\begin{aligned} &\mathrm{PQ100}=D_v^2G(c,a_0)+D_v^2G(\bar{c},a_2)+\tfrac{1}{4}D_v^3G(c,c,\bar{c}), \\ &\mathrm{PQ010}=D_vD_\lambda G(c), \end{aligned} \tag{3.10}$$

before computing

$$p_{100}=\tfrac{1}{4}\,\mathrm{Real}\,(d^*\mathrm{PQ100}), \qquad p_{010}=\tfrac{1}{2}\,\mathrm{Real}\,(d^*\mathrm{PQ010}). \tag{3.11}$$

The coefficients q_{100} and q_{010} are obtained by taking imaginary instead of real parts and inserting a minus sign. It is a simple matter to identify p_{010} with $\sigma_\lambda(0,0)$, so that a nonzero value means the transversality condition is satisfied. If p_{100} is not zero, then a negative value indicates a stable periodic orbit and vice versa.

Higher-order calculations become increasingly complex. We need eight more vectors to compute the second-order coefficients; six are obtained by solving:

(3.12)

$$\begin{aligned} &(A-3i\omega_0 I)a_3=-\tfrac{3}{2}D_v^2G(c,a_2)-\tfrac{1}{8}D_v^3G(c,c,c), \\ &Ab_0=-2[D_v^2G(c,\bar{a}_1)+D_v^2G(\bar{c},a_1)]-3D_v^2G(a_0,a_0)-6D_v^2G(a_2,\bar{a}_2) \\ &\qquad -3D_v^3G(c,\bar{c},a_0)-\tfrac{3}{2}D_v^3G(c,c,\bar{a}_2)-\tfrac{3}{2}D_v^3G(\bar{c},\bar{c},a_2)-\tfrac{3}{8}D_v^4G(c,c,\bar{c},\bar{c}), \\ &(A-2i\omega_0 I)b_2=-2D_v^2G(c,a_1)-2D_v^2G(\bar{c},a_3)-6D_v^2G(a_0,a_2) \\ &\qquad -3D_v^3G(c,\bar{c},a_2)-\tfrac{3}{2}D_v^3G(c,c,a_0)-\tfrac{1}{4}D_v^4G(c,c,c,\bar{c}), \\ &(A-2i\omega_0 I)c_2=2i\omega_0 a_2, \\ &Ad_0=-\tfrac{1}{2}D_v^2G(c,\bar{c}_1)-\tfrac{1}{2}D_v^2G(\bar{c},c_1)-\tfrac{1}{2}D_v^2D_\lambda G(c,\bar{c})-D_vD_\lambda G(a_0) \\ &(A-2i\omega_0 I)d_2=-\tfrac{1}{2}D_v^2G(c,c_1)-\tfrac{1}{4}D_v^2D_\lambda G(c,c)-D_vD_\lambda G(a_2). \end{aligned}$$

The remaining two vectors require some explanation, since the left-hand sides of the equations are singular. According to the Fredholm alternative, the solution exists and is unique if we specify that the solution be orthogonal to the eigenvector of the adjoint problem and if the right-hand side of the equation is orthogonal to the same eigenvector. This eigenvector d is found from

$$(A^* + i\omega_0 I)d = 0 \tag{3.13}$$

(where A^* is the transpose of A) and normalized so that $d^*c = 2$. The two remaining vectors thus are uniquely determined from

$$\begin{aligned} &(A - i\omega_0 I)a_1 = -\tfrac{3}{2}\mathrm{PQ100} + \tfrac{3}{4}[d^*\mathrm{PQ100}]c, \qquad d^*a_1 = 0, \\ &(A - i\omega_0 I)c_1 = -\mathrm{PQ010} + \tfrac{1}{2}[d^*\mathrm{PQ010}]c, \qquad d^*c_1 = 0. \end{aligned} \tag{3.14}$$

The next step in determining the second-order coefficients is to compute the following PQijk quantities, which will lead directly to the p_{ijk} and q_{ijk} quantities we desire:

(3.15)

$$\mathrm{PQ101} = D_v^2 G(\bar{c}, c_2), \qquad \mathrm{PQ020} = D_v D_\lambda G(c_1) + \tfrac{1}{2} D_v D_\lambda^2 G(c),$$

$$\begin{aligned} \mathrm{PQ110} = {} & D_v^2 G(c_1, a_0) + D_v^2 G(\bar{c}_1, a_2) + D_v^2 G(c, d_0) + D_v^2 G(\bar{c}, d_2) + D_v^2 D_\lambda G(c, a_0) \\ & + D_v^2 D_\lambda G(\bar{c}, a_2) + \tfrac{2}{3} D_v D_\lambda G(a_1) + \tfrac{1}{4} D_v^3 G(c, c, \bar{c}_1) \\ & + \tfrac{1}{2} D_v^3 G(c, \bar{c}, c_1) + \tfrac{1}{4} D_v^3 D_\lambda G(c, c, \bar{c}) \end{aligned}$$

$$\begin{aligned} \mathrm{PQ200} = {} & \tfrac{1}{2} D_v^2 G(c, b_0) + \tfrac{1}{2} D_v^2 G(\bar{c}, b_2) + 2 D_v^2 G(a_0, a_1) + 2 D_v^2 G(a_2, \bar{a}_1) + 2 D_v^2 G(\bar{a}_2, a_3) \\ & + \tfrac{1}{2} D_v^3 G(c, c, \bar{a}_1) + D_v^3 G(c, \bar{c}, a_1) + \tfrac{1}{2} D_v^3 G(\bar{c}, \bar{c}, a_3) \\ & + 3 D_v^3 G(c, a_2, \bar{a}_2) + 3 D_v^3 G(\bar{c}, a_0, a_2) + \tfrac{3}{2} D_v^3 G(c, a_0, a_0) \\ & + \tfrac{1}{4} D_v^4 G(c, c, c, \bar{a}_2) + \tfrac{3}{4} D_v^4 G(c, c, \bar{c}, a_0) + \tfrac{3}{4} D_v^4 G(c, \bar{c}, \bar{c}, a_2) \\ & + \tfrac{1}{16} D_v^5 G(c, c, c, \bar{c}, \bar{c}). \end{aligned}$$

The p_{ijk} coefficients for $i+j+k=2$ are now obtained from the formulas

$$\begin{aligned} &p_{101} = \tfrac{1}{4}\,\mathrm{Real}\,(d^*\mathrm{PQ101}), \qquad p_{020} = \tfrac{1}{2}\,\mathrm{Real}\,(d^*\mathrm{PQ020}), \\ &p_{110} = \tfrac{1}{4}\,\mathrm{Real}\,(d^*\mathrm{PQ110}), \qquad p_{200} = \tfrac{1}{24}\,\mathrm{Real}\,(d^*\mathrm{PQ200}), \end{aligned} \tag{3.16}$$

and the analogous q_{ijk} coefficients, by changing the sign and taking the imaginary instead of the real part.

Computing the third-order coefficients is even more complicated, but proceeds in exactly the same way. We first compute 24 vectors; these vectors, along with the ten already computed, could be used to obtain approximations to the periodic orbits, but we have not done this. We will only use them to obtain the p_{ijk} coefficients for $i+j+k=3$. As far as we know, these coefficients have not been obtained before.

The first set of vectors are required for the PQ300 quantity and are obtained by solving the following series of problems:

(3.17)

$$(A - 4i\omega_0 I)b_4 = -3D_v^2 G(a_2, a_2) - 2D_v^2 G(c, a_3) - \tfrac{3}{2} D_v^3 G(c, c, a_2) - \tfrac{1}{16} D_v^4 G(c, c, c, c),$$

$$(A - i\omega_0 I)e_1 = -5\mathrm{PQ200} + \tfrac{5}{2}[d^*\mathrm{PQ200}]c, \qquad d^*e_1 = 0,$$

$$(A-3i\omega_0 I)e_3 = -\tfrac{5}{2}D_v^2 G(\bar{c}, b_4) - \tfrac{5}{2}D_v^2 G(c, b_2) - 10D_v^2 G(a_0, a_3) - 10D_v^2 G(a_2, a_1)$$
$$- 5D_v^3 G(c, \bar{c}, a_3) - \tfrac{5}{2}D_v^3 G(c, c, a_1) - 15D_v^3 G(c, a_0, a_2)$$
$$- \tfrac{15}{2}D_v^3 G(\bar{c}, a_2, a_2)$$
$$- \tfrac{15}{4}D_v^4 G(c, c, \bar{c}, a_2) - \tfrac{5}{4}D_v^4 G(c, c, c, a_0) - \tfrac{5}{32}D_v^5 G(c, c, c, c, \bar{c}),$$

$$Ah_0 = -3D_v^2 G(\bar{c}, e_1) - 3D_v^2 G(c, \bar{e}_1) - 15D_v^2 G(a_2, \bar{b}_2) - 15D_v^2 G(\bar{a}_2, b_2)$$
$$- 15D_v^2 G(a_0, b_0)$$
$$- 20D_v^2 G(a_1, \bar{a}_1) - 20D_v^2 G(a_3, \bar{a}_3) - \tfrac{15}{4}D_v^3 G(\bar{c}, \bar{c}, b_2) - \tfrac{15}{4}D_v^3 G(c, c, \bar{b}_2)$$
$$- \tfrac{15}{2}D_v^3 G(c, \bar{c}, b_0) - 30D_v^3 G(\bar{c}, \bar{a}_2, a_3) - 30D_v^3 G(c, a_2, \bar{a}_3)$$
$$- 30D_v^3 G(\bar{c}, a_0, a_1)$$
$$- 30D_v^3 G(c, a_0, \bar{a}_1) - 30D_v^3 G(c, \bar{a}_2, a_1) - 30D_v^3 G(\bar{c}, a_2, \bar{a}_1)$$
$$- 15D_v^3 G(a_0, a_0, a_0)$$
$$- 90D_v^3 G(a_2, \bar{a}_2, a_0) - \tfrac{5}{2}D_v^4 G(\bar{c}, \bar{c}, \bar{c}, a_3) - \tfrac{5}{2}D_v^4 G(c, c, c, \bar{a}_3)$$
$$- \tfrac{15}{2}D_v^4 G(c, \bar{c}, \bar{c}, a_1)$$
$$- \tfrac{15}{2}D_v^4 G(\bar{c}, c, c, \bar{a}_1) - \tfrac{45}{2}D_v^4 G(\bar{c}, \bar{c}, a_2, a_0) - \tfrac{45}{2}D_v^4 G(c, c, \bar{a}_2, a_0)$$
$$- 45D_v^4 G(c, \bar{c}, a_2, \bar{a}_2)$$
$$- \tfrac{45}{2}D_v^4 G(c, \bar{c}, a_0, a_0) - \tfrac{15}{4}D_v^5 G(c, \bar{c}, \bar{c}, \bar{c}, a_2) - \tfrac{15}{4}D_v^5 G(c, c, c, \bar{c}, \bar{a}_2)$$
$$- \tfrac{45}{8}D_v^5 G(c, c, \bar{c}, \bar{c}, a_0) - \tfrac{5}{16}D_v^6 G(c, c, c, \bar{c}, \bar{c}, \bar{c}),$$

$$(A-2i\omega_0 I)h_2 = -3D_v^2 G(\bar{c}, e_3) - 3D_v^2 G(c, e_1) - 15D_v^2 G(\bar{a}_2, b_4) - 15D_v^2 G(a_0, b_2)$$
$$- 15D_v^2 G(a_2, b_0) - 10D_v^2 G(a_1, a_1) - 20D_v^2 G(a_3, \bar{a}_1)$$
$$- \tfrac{15}{4}D_v^3 G(\bar{c}, \bar{c}, b_4) - \tfrac{15}{2}D_v^3 G(c, \bar{c}, b_2)$$
$$- \tfrac{15}{4}D_v^3 G(c, c, b_0) - 30D_v^3 G(c, \bar{a}_2, a_3) - 30D_v^3 G(\bar{c}, a_0, a_3)$$
$$- 30D_v^3 G(\bar{c}, a_2, a_1) - 30D_v^3 G(c, a_0, a_1) - 30D_v^3 G(c, a_2, \bar{a}_1)$$
$$- 45D_v^3 G(a_2, a_2, \bar{a}_2)$$
$$- 45D_v^3 G(a_2, a_0, a_0) - \tfrac{15}{2}D_v^4 G(c, \bar{c}, \bar{c}, a_3) - \tfrac{5}{2}D_v^4 G(c, c, c, \bar{a}_1)$$
$$- \tfrac{15}{2}D_v^4 G(c, c, \bar{c}, a_1)$$
$$- \tfrac{45}{4}D_v^4 G(\bar{c}, \bar{c}, a_2, a_2) - 45D_v^4 G(c, \bar{c}, a_2, a_0) - \tfrac{45}{2}D_v^4 G(c, c, a_2, \bar{a}_2)$$
$$- \tfrac{45}{4}D_v^4 G(c, c, a_0, a_0)$$
$$- \tfrac{45}{8}D_v^5 G(c, c, \bar{c}, \bar{c}, a_2) - \tfrac{15}{4}D_v^5 G(c, c, c, \bar{c}, a_0) - \tfrac{15}{16}D_v^5 G(c, c, c, c, \bar{a}_2)$$
$$- \tfrac{15}{64}D_v^6 G(c, c, c, c, \bar{c}, \bar{c}).$$

PQ210 vectors:

(3.18)

$$
\begin{aligned}
An_0 = &-D_v D_\lambda G(b_0) - 2D_v^2 D_\lambda G(\bar{c}, a_1) - 2D_v^2 D_\lambda G(c, \bar{a}_1) - 3D_v^2 D_\lambda G(a_0, a_0) \\
&- 6D_v^2 D_\lambda G(a_2, \bar{a}_2) \\
&- \tfrac{3}{2} D_v^3 D_\lambda G(\bar{c}, \bar{c}, a_2) - \tfrac{3}{2} D_v^3 D_\lambda G(c, c, \bar{a}_2) - 3D_v^3 D_\lambda G(c, \bar{c}, a_0) \\
&- \tfrac{3}{8} D_v^4 D_\lambda G(c, c, \bar{c}, \bar{c}) \\
&- 2D_v^2 G(\bar{c}, n_1) - 2D_v^2 G(c, \bar{n}_1) - 6D_v^2 G(\bar{a}_2, d_2) - 6D_v^2 G(a_2, \bar{d}_2) \\
&- 6D_v^2 G(a_0, d_0) - 2D_v^2 G(a_1, \bar{c}_1) - 2D_v^2 G(\bar{a}_1, c_1) - \tfrac{3}{2} D_v^3 G(c, c, \bar{d}_2) \\
&- \tfrac{3}{2} D_v^3 G(\bar{c}, \bar{c}, d_2) - 3D_v^3 G(c, \bar{c}, d_0) - 3D_v^3 G(\bar{c}, \bar{c}_1, a_2) - 3D_v^3 G(c, c_1, \bar{a}_2) \\
&- 3D_v^3 G(\bar{c}, c_1, a_0) - 3D_v^3 G(c, \bar{c}_1, a_0) - \tfrac{3}{4} D_v^4 G(\bar{c}, \bar{c}, c, c_1) - \tfrac{3}{4} D_v^4 G(c, c, \bar{c}, \bar{c}_1),
\end{aligned}
$$

$$
(A - i\omega_0 I) n_1 = -\tfrac{3}{2}\mathrm{PQ110} + \tfrac{3}{4}[d^*\mathrm{PQ110}]c, \qquad d^* n_1 = 0,
$$

$$
\begin{aligned}
(A - 2i\omega_0 I) n_2 = &-D_v D_\lambda G(b_2) - 2D_v^2 D_\lambda G(\bar{c}, a_3) - 2D_v^2 D_\lambda G(c, a_1) - 6D_v^2 D_\lambda G(a_2, a_0) \\
&- 3D_v^3 D_\lambda G(c, \bar{c}, a_2) - \tfrac{3}{2} D_v^3 D_\lambda G(c, c, a_0) - \tfrac{1}{4} D_v^4 D_\lambda G(c, c, c, \bar{c}) \\
&- 2D_v^2 G(\bar{c}, n_3) \\
&- 2D_v^2 G(c, n_1) - 6D_v^2 G(a_0, d_2) - 6D_v^2 G(a_2, d_0) - 2D_v^2 G(a_1, c_1) \\
&- 2D_v^2 G(a_3, \bar{c}_1) - 3D_v^3 G(c, \bar{c}, d_2) - \tfrac{3}{2} D_v^3 G(c, c, d_0) \\
&- 3D_v^3 G(\bar{c}, c_1, a_2) \\
&- 3D_v^3 G(c, \bar{c}_1, a_2) - 3D_v^3 G(c, c_1, a_0) - \tfrac{3}{4} D_v^4 G(c, c, \bar{c}, c_1) \\
&- \tfrac{1}{4} D_v^4 G(c, c, c, \bar{c}_1),
\end{aligned}
$$

$$
\begin{aligned}
(A - 3i\omega_0 I) n_3 = &-D_v D_\lambda G(a_3) - \tfrac{3}{2} D_v^2 D_\lambda G(c, a_2) - \tfrac{1}{8} D_v^3 D_\lambda G(c, c, c) - \tfrac{3}{2} D_v^2 G(c, d_2) \\
&- \tfrac{3}{2} D_v^2 G(a_2, c_1) - \tfrac{3}{8} D_v^3 G(c, c, c_1).
\end{aligned}
$$

PQ201 vectors:

(3.19)

$$
\begin{aligned}
Ak_0 = &-2D_v^2 G(\bar{c}, k_1) - 2D_v^2 G(c, \bar{k}_1) - 6D_v^2 G(\bar{a}_2, c_2) - 6D_v^2 G(a_2, \bar{c}_2) \\
&- \tfrac{3}{2} D_v^3 G(\bar{c}, \bar{c}, c_2) - \tfrac{3}{2} D_v^3 G(c, c, \bar{c}_2),
\end{aligned}
$$

$$
(A - i\omega_0 I) k_1 = -\tfrac{3}{2}\mathrm{PQ101} + \tfrac{3}{4}[d^*\mathrm{PQ101}]c + i\omega_0 a_1, \qquad d^* k_1 = 0,
$$

$$
(A - 2i\omega_0 I) k_2 = 2i\omega_0 b_2 - 2D_v^2 G(\bar{c}, k_3) - 2D_v^2 G(c, k_1) - 6D_v^2 G(a_0, c_2) - 3D_v^3 G(c, \bar{c}, c_2),
$$

$$
(A - 3i\omega_0 I) k_3 = 3i\omega_0 a_3 - \tfrac{3}{2} D_v^2 G(c, c_2).
$$

PQ111 vectors:

(3.20)

$$
\begin{aligned}
&Aj_0 = -D_v^2 G(c, \bar{j}_1) - D_v^2 G(\bar{c}, j_1), \\
&(A - i\omega_0 I) j_1 = \tfrac{1}{2} i\omega_0 c_1, \qquad d^* j_1 = 0, \\
&(A - 2i\omega_0 I) j_2 = 2i\omega_0 d_2 - D_v^2 G(c, j_1) - D_v D_\lambda G(c_2).
\end{aligned}
$$

PQ102 vector:

(3.21)

$$
(A - 2i\omega_0 I) m_2 = 4i\omega_0 c_2.
$$

PQ120 vectors:

(3.22)

$$Ar_0 = -D_v^2 G(c_1, \bar{c}_1) - D_v^2 G(c, \bar{r}_1) - D_v^2 G(\bar{c}, r_1) - 2D_v D_\lambda G(d_0) - D_v^2 D_\lambda G(\bar{c}, c_1)$$
$$- D_v^2 D_\lambda G(c, \bar{c}_1) - D_v D_\lambda^2 G(a_0) - \tfrac{1}{2} D_v^2 D_\lambda^2 G(c, \bar{c}),$$
$$(A - i\omega_0 I) r_1 = -\mathrm{PQ020} + \tfrac{1}{2}[d^* \mathrm{PQ020}]c, \qquad d^* r_1 = 0,$$
$$(A - 2i\omega_0 I) r_2 = -D_v^2 G(c, r_1) - \tfrac{1}{2} D_v^2 G(c_1, c_1) - 2D_v D_\lambda G(d_2) - D_v^2 D_\lambda G(c, c_1)$$
$$- D_v D_\lambda^2 G(a_2) - \tfrac{1}{4} D_v^2 D_\lambda^2 G(c, c).$$

Next we define analogous PQijk quantities for $i+j+k=3$.

(3.23)

$$\mathrm{PQ300} = \tfrac{1}{2} D_v^2 G(\bar{c}, h_2) + \tfrac{1}{2} D_v^2 G(c, h_0) + 3D_v^2 G(\bar{a}_2, e_3) + 3D_v^2 G(a_0, e_1) + 3D_v^2 G(a_2, \bar{e}_1)$$
$$+ 5D_v^2 G(\bar{a}_3, b_4) + 5D_v^2 G(\bar{a}_1, b_2) + 5D_v^2 G(a_1, b_0) + 5D_v^2 G(a_3, \bar{b}_2)$$
$$+ \tfrac{3}{4} D_v^3 G(\bar{c}, \bar{c}, e_3) + \tfrac{3}{2} D_v^3 G(c, \bar{c}, e_1) + \tfrac{3}{4} D_v^3 G(c, c, \bar{e}_1) + \tfrac{15}{2} D_v^3 G(\bar{c}, a_0, b_2)$$
$$+ \tfrac{15}{2} D_v^3 G(\bar{c}, \bar{a}_2, b_4) + \tfrac{15}{2} D_v^3 G(c, \bar{a}_2, b_2) + \tfrac{15}{2} D_v^3 G(\bar{c}, a_2, b_0)$$
$$+ \tfrac{15}{2} D_v^3 G(c, a_0, b_0)$$
$$+ \tfrac{15}{2} D_v^3 G(c, a_2, \bar{b}_2) + 5D_v^3 G(\bar{c}, a_1, a_1) + 10D_v^3 G(\bar{c}, a_3, \bar{a}_1)$$
$$+ 10D_v^3 G(c, a_1, \bar{a}_1)$$
$$+ 10D_v^3 G(c, a_3, \bar{a}_3) + 30D_v^3 G(a_0, \bar{a}_2, a_3) + 30D_v^3 G(a_2, \bar{a}_2, a_1)$$
$$+ 15D_v^3 G(a_0, a_0, a_1)$$
$$+ 30D_v^3 G(a_0, a_2, \bar{a}_1) + 15D_v^3 G(a_2, a_2, \bar{a}_3) + \tfrac{5}{8} D_v^4 G(\bar{c}, \bar{c}, \bar{c}, b_4)$$
$$+ \tfrac{15}{8} D_v^4 G(c, \bar{c}, \bar{c}, b_2)$$
$$+ \tfrac{15}{8} D_v^4 G(c, c, \bar{c}, b_0) + \tfrac{5}{8} D_v^4 G(c, c, c, \bar{b}_2) + 15D_v^4 G(c, \bar{c}, \bar{a}_2, a_3)$$
$$+ \tfrac{15}{2} D_v^4 G(\bar{c}, \bar{c}, a_0, a_3)$$
$$+ \tfrac{15}{2} D_v^4 G(\bar{c}, \bar{c}, a_2, a_1) + 15D_v^4 G(c, \bar{c}, a_0, a_1) + \tfrac{15}{2} D_v^4 G(c, c, \bar{a}_2, a_1)$$
$$+ 15D_v^4 G(c, \bar{c}, a_2, \bar{a}_1)$$
$$+ \tfrac{15}{2} D_v^4 G(c, c, a_0, \bar{a}_1) + \tfrac{15}{2} D_v^4 G(c, c, a_2, \bar{a}_3) + \tfrac{15}{2} D_v^4 G(c, a_0, a_0, a_0)$$
$$+ \tfrac{45}{2} D_v^4 G(\bar{c}, a_0, a_0, a_2)$$
$$+ \tfrac{45}{2} D_v^4 G(\bar{c}, a_2, a_2, \bar{a}_2) + 45D_v^4 G(c, a_0, a_2, \bar{a}_2) + \tfrac{5}{4} D_v^5 G(c, \bar{c}, \bar{c}, \bar{c}, a_3)$$
$$+ \tfrac{15}{8} D_v^5 G(c, c, \bar{c}, \bar{c}, a_1)$$
$$+ \tfrac{5}{4} D_v^5 G(c, c, c, \bar{c}, \bar{a}_1) + \tfrac{5}{16} D_v^5 G(c, c, c, c, \bar{a}_3) + \tfrac{15}{8} D_v^5 G(\bar{c}, \bar{c}, \bar{c}, a_2, a_2)$$
$$+ \tfrac{45}{4} D_v^5 G(c, \bar{c}, \bar{c}, a_0, a_2)$$
$$+ \tfrac{45}{8} D_v^5 G(c, c, \bar{c}, a_0, a_0) + \tfrac{45}{4} D_v^5 G(c, c, \bar{c}, a_2, \bar{a}_2) + \tfrac{15}{4} D_v^5 G(c, c, c, a_0, \bar{a}_2)$$
$$+ \tfrac{15}{16} D_v^6 G(c, c, \bar{c}, \bar{c}, \bar{c}, a_2)$$
$$+ \tfrac{15}{16} D_v^6 G(c, c, c, \bar{c}, \bar{c}, a_0) + \tfrac{15}{32} D_v^6 G(c, c, c, c, \bar{c}, \bar{a}_2)$$
$$+ \tfrac{5}{128} D_v^7 G(c, c, c, c, \bar{c}, \bar{c}, \bar{c});$$

(3.24)

$$\begin{aligned}\mathrm{PQ210} = {} & \tfrac{1}{5} D_v D_\lambda G(e_1) + \tfrac{1}{2} D_v^2 D_\lambda G(c, b_0) + \tfrac{1}{2} D_v^2 D_\lambda G(\bar{c}, b_2) + 2 D_v^2 D_\lambda G(a_2, \bar{a}_1) \\ & + 2 D_v^2 D_\lambda G(a_0, a_1) + 2 D_v^2 D_\lambda G(\bar{a}_2, a_3) + \tfrac{1}{2} D_v^3 D_\lambda G(c, c, \bar{a}_1) \\ & + D_v^3 D_\lambda G(c, \bar{c}, a_1) \\ & + \tfrac{1}{2} D_v^3 D_\lambda G(\bar{c}, \bar{c}, a_3) + 3 D_v^3 D_\lambda G(c, a_2, \bar{a}_2) + \tfrac{3}{2} D_v^3 D_\lambda G(c, a_0, a_0) \\ & + 3 D_v^3 D_\lambda G(\bar{c}, a_2, a_0) \\ & + \tfrac{1}{4} D_v^4 D_\lambda G(c, c, c, \bar{a}_2) + \tfrac{3}{4} D_v^4 D_\lambda G(c, c, \bar{c}, a_0) + \tfrac{3}{4} D_v^4 D_\lambda G(c, \bar{c}, \bar{c}, a_2) \\ & + \tfrac{1}{16} D_v^5 D_\lambda G(c, c, c, \bar{c}, \bar{c}) \\ & + \tfrac{1}{2} D_v^2 G(c, n_0) + \tfrac{1}{2} D_v^2 G(\bar{c}, n_2) + 2 D_v^2 G(a_2, \bar{n}_1) + 2 D_v^2 G(a_0, n_1) \\ & + 2 D_v^2 G(\bar{a}_2, n_3) + 2 D_v^2 G(a_3, \bar{d}_2) + 2 D_v^2 G(a_1, d_0) + 2 D_v^2 G(\bar{a}_1, d_2) \\ & + \tfrac{1}{2} D_v^2 G(b_2, \bar{c}_1) + \tfrac{1}{2} D_v^2 G(b_0, c_1) + \tfrac{1}{2} D_v^3 G(c, c, \bar{n}_1) + D_v^3 G(c, \bar{c}, n_1) \\ & + \tfrac{1}{2} D_v^3 G(\bar{c}, \bar{c}, n_3) + 3 D_v^3 G(c, a_2, \bar{d}_2) + 3 D_v^3 G(c, a_0, d_0) + 3 D_v^3 G(c, \bar{a}_2, d_2) \\ & + 3 D_v^3 G(\bar{c}, a_2, d_0) + 3 D_v^3 G(\bar{c}, a_0, d_2) + D_v^3 G(c, a_1, \bar{c}_1) + D_v^3 G(c, \bar{a}_1, c_1) \\ & + D_v^3 G(\bar{c}, a_3, \bar{c}_1) + D_v^3 G(\bar{c}, a_1, c_1) + 3 D_v^3 G(a_2, a_0, \bar{c}_1) \\ & + 3 D_v^3 G(a_2, \bar{a}_2, c_1) \\ & + \tfrac{3}{2} D_v^3 G(a_0, a_0, c_1) + \tfrac{1}{4} D_v^4 G(c, c, c, \bar{d}_2) + \tfrac{3}{4} D_v^4 G(c, c, \bar{c}, d_0) \\ & + \tfrac{3}{4} D_v^4 G(c, \bar{c}, \bar{c}, d_2) \\ & + \tfrac{3}{4} D_v^4 G(c, c, a_0, \bar{c}_1) + \tfrac{3}{4} D_v^4 G(c, c, \bar{a}_2, c_1) + \tfrac{3}{2} D_v^4 G(c, \bar{c}, a_2, \bar{c}_1) \\ & + \tfrac{3}{2} D_v^4 G(c, \bar{c}, a_0, c_1) \\ & + \tfrac{3}{4} D_v^4 G(\bar{c}, \bar{c}, a_2, c_1) + \tfrac{1}{8} D_v^5 G(c, c, c, \bar{c}, \bar{c}_1) + \tfrac{3}{16} D_v^5 G(c, c, \bar{c}, \bar{c}, c_1); \end{aligned}$$

(3.25)

$$\begin{aligned}\mathrm{PQ201} = {} & \tfrac{1}{2} D_v^2 G(\bar{c}, k_2) + \tfrac{1}{2} D_v^2 G(c, k_0) + 2 D_v^2 G(\bar{a}_2, k_3) + 2 D_v^2 G(a_0, k_1) + 2 D_v^2 G(a_2, \bar{k}_1) \\ & + 2 D_v^2 G(\bar{a}_1, c_2) + 2 D_v^2 G(a_3, \bar{c}_2) + \tfrac{1}{2} D_v^3 G(\bar{c}, \bar{c}, k_3) + D_v^3 G(c, \bar{c}, k_1) \\ & + \tfrac{1}{2} D_v^3 G(c, c, \bar{k}_1) + 3 D_v^3 G(c, \bar{c}_2, a_2) + 3 D_v^3 G(\bar{c}, c_2, a_0) + 3 D_v^3 G(c, c_2, \bar{a}_2) \\ & + \tfrac{3}{4} D_v^4 G(c, \bar{c}, \bar{c}, c_2) + \tfrac{1}{4} D_v^4 G(c, c, c, \bar{c}_2); \end{aligned}$$

(3.26)

$$\begin{aligned}\mathrm{PQ111} = {} & \tfrac{1}{3} D_v D_\lambda G(k_1) + \tfrac{1}{2} D_v^2 D_\lambda G(\bar{c}, c_2) + \tfrac{1}{2} D_v^2 G(\bar{c}_1, c_2) \\ & + D_v^2 G(a_2, \bar{j}_1) + D_v^2 G(a_0, j_1) + \tfrac{1}{2} D_v^2 G(c, j_0) \\ & + \tfrac{1}{2} D_v^2 G(\bar{c}, j_2) + \tfrac{1}{4} D_v^3 G(c, c, \bar{j}_1) + \tfrac{1}{2} D_v^3 G(c, \bar{c}, j_1); \end{aligned}$$

(3.27)

$$\mathrm{PQ102} = \tfrac{1}{2} D_v^2 G(\bar{c}, m_2);$$

(3.28)

$$\begin{aligned}\mathrm{PQ120} = {} & \tfrac{1}{2} D_v^2 G(\bar{c}, r_2) + \tfrac{1}{2} D_v^2 G(c, r_0) + D_v^2 G(a_0, r_1) + D_v^2 G(a_2, \bar{r}_1) + D_v^2 G(d_0, c_1) \\ & + D_v^2 G(d_2, \bar{c}_1) + \tfrac{1}{4} D_v^3 G(c, c, \bar{r}_1) + \tfrac{1}{2} D_v^3 G(c, \bar{c}, r_1) + \tfrac{1}{4} D_v^3 G(\bar{c}, c_1, c_1) \\ & + \tfrac{1}{2} D_v^3 G(c, c_1, \bar{c}_1) + \tfrac{2}{3} D_v D_\lambda G(n_1) + D_v^2 D_\lambda G(\bar{c}, d_2) + D_v^2 D_\lambda G(c, d_0) \\ & + D_v^2 D_\lambda G(a_0, c_1) + D_v^2 D_\lambda G(a_2, \bar{c}_1) + \tfrac{1}{2} D_v^3 D_\lambda G(c, \bar{c}, c_1) \\ & + \tfrac{1}{4} D_v^3 D_\lambda G(c, c, \bar{c}_1) \\ & + \tfrac{1}{3} D_v D_\lambda^2 G(a_1) + \tfrac{1}{2} D_v^2 D_\lambda^2 G(\bar{c}, a_2) + \tfrac{1}{2} D_v^2 D_\lambda^2 G(c, a_0) + \tfrac{1}{8} D_v^3 D_\lambda^2 G(c, c, \bar{c}); \end{aligned}$$

$$\text{PQ021} = D_v D_\lambda G(j_1); \tag{3.29}$$

$$\text{PQ030} = D_v D_\lambda G(r_1) + \tfrac{1}{2} D_v D_\lambda^2 G(c_1) + \tfrac{1}{6} D_v D_\lambda^3 G(c). \tag{3.30}$$

Finally we obtain the p_{ijk} coefficients from

$$\begin{aligned} p_{300} &= \text{Real}\,(d^*\text{PQ300})/6!, & p_{210} &= \text{Real}\,(d^*\text{PQ210})/4!, \\ p_{201} &= \text{Real}\,(d^*\text{PQ201})/4!, & p_{111} &= \text{Real}\,(d^*\text{PQ111})/2, \\ p_{102} &= \text{Real}\,(d^*\text{PQ102})/4, & p_{120} &= \text{Real}\,(d^*\text{PQ120})/4, \\ p_{021} &= \text{Real}\,(d^*\text{PQ021}), & p_{030} &= \text{Real}\,(d^*\text{PQ030})/2, \end{aligned} \tag{3.31}$$

and the analogous q_{ijk} coefficients can be obtained by changing the sign and taking imaginary parts instead of real parts.

4. Application to Hilbert's 16th problem. There is a large literature on Hilbert's 16th problem; the reader is referred to the review articles of Coppel [9], [10] and Ye [32], [33]. As remarked in the Introduction, the maximum number of limit cycles of a quadratic system in the plane is still not known. The general quadratic system is

$$\frac{dx}{dt} = \sum_{i+j\leqq 2} A_{ij} x^i y^j, \qquad \frac{dy}{dt} = \sum_{i+j\leqq 2} B_{ij} x^i y^j. \tag{4.1}$$

Let $H_2(A, B)$ denote the total number of limit cycles of (4.1) for given A_{ij} and B_{ij}. Then it is known that

$$H_2(A, B) < \infty, \qquad \sup_{A,B} H_2(A, B) \geqq 4. \tag{4.2}$$

It was shown by Bamón [3] that $H_2(A, B)$ is finite, after several false proofs dating back to Dulac. Examples of (4.1) with 4 limit cycles were constructed only recently by Shi [30] and Chen and Wang [7]. Both of these examples employ Theorem 2.1, or what is known as Bautin's technique: perturbation of a weak focus to create multiple limit cycles locally. Shi's example has a weak focus of order 3 and a strong focus surrounded by a unique limit cycle; Wang's example has a weak focus of order 2 and a strong focus, each of which is surrounded by a finite limit cycle. These local arguments gain added significance in view of the following result.

THEOREM (Li [24]). *There is no limit cycle around a weak focus of order* 3, *for any quadratic system.*

Thus it is not possible to construct an example of a quadratic system with 4 limit cycles about the same equilibrium, using Bautin's technique.

The remainder of this section is devoted to a new derivation of Bautin's formulas [4]. First we restate Bautin's results. Assuming that (4.1) has an equilibrium that is a focus, Bautin makes a linear transformation to the system

$$\begin{aligned} x' &= -y + \lambda_1 x - \lambda_3 x^2 + (2\lambda_2 + \lambda_5) xy + \lambda_6 y^2, \\ y' &= x + \lambda_1 y + \lambda_2 x^2 + (2\lambda_3 + \lambda_4) xy - \lambda_2 y^2. \end{aligned} \tag{4.3}$$

Then the focal value $c_0(0)$ defined in § 2 is clearly equal to λ_1, and corresponding to the focal values $c_2(0)$, $c_4(0)$, and $c_6(0)$, Bautin calculated (by the method of the succession function) the following quantities:

$$\begin{aligned} v_3 &= -(\pi/4)\lambda_5(\lambda_3 - \lambda_6), \\ v_5 &= (\pi/24)\lambda_2\lambda_4(\lambda_3 - \lambda_6)[\lambda_4 + 5(\lambda_3 - \lambda_6)], \\ v_7 &= (25\pi/32)\lambda_2\lambda_4(\lambda_3 - \lambda_6)^2[\lambda_3\lambda_6 - 2\lambda_6^2 - \lambda_2^2]. \end{aligned} \tag{4.4}$$

The origin for (4.3) is a weak focus of order k, for $1 \leqq k \leqq 3$, if

$$\lambda_1 = \cdots = v_{2k-1} = 0, \qquad v_{2k+1} \neq 0. \tag{4.5}$$

Moreover, Bautin showed that the origin is a center, if and only if

$$\lambda_1 = v_3 = v_5 = v_7 = 0. \tag{4.6}$$

(Recently Li [23] has reformulated Bautin's results, in terms of the coefficients of the untransformed equation (4.1), for greater convenience in application.)

Now let us apply the formulas of § 3 to (4.3), with $\lambda_1 = 0$. Note that there exists a trivial solution $(x, y) = (0, 0)$, the frequency is scaled to be 1, and all of the derivatives higher than second-order are zero, so the calculation is greatly simplified. We find

$$\begin{gathered} A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad c = d = \begin{pmatrix} i \\ 1 \end{pmatrix}, \\ a_0 = \begin{pmatrix} 0 \\ -(\lambda_3 - \lambda_6) \end{pmatrix}, \qquad a_2 = (1/6)\begin{pmatrix} 2(\lambda_2 + \lambda_5) + i(\lambda_4 - 2\lambda_6) \\ 3\lambda_3 + 2\lambda_4 - \lambda_6 + i(2\lambda_2 - \lambda_5) \end{pmatrix}. \end{gathered} \tag{4.7}$$

From this we obtain the first focal value

$$\begin{aligned} a_{10} = p_{100} &= \tfrac{1}{4}\,\mathrm{Real}\,[d^*(D_u^2 f(c, a_0) + D_u^2 f(c, a_2))] \\ &= -\tfrac{1}{8}\lambda_5(\lambda_3 - \lambda_6), \end{aligned} \tag{4.8}$$

in agreement with Bautin's result, up to a scaling factor of 2π.

Before proceeding, we make a simplifying assumption. The second focal value is of interest only if the first focal value is 0. Therefore we may assume that $a_{10} = 0$ in (4.8). It can be shown directly that the origin is a center for (4.3) if

$$\lambda_1 = (\lambda_3 - \lambda_6) = 0. \tag{4.9}$$

Therefore the only possibility for the origin to be a weak focus of order 2 for (4.3) is if

$$\lambda_1 = \lambda_5 = 0, \tag{4.10}$$

which we assume henceforth. Then for the second focal value the formulas in § 3 give

$$a_{20} = \tfrac{1}{48}\lambda_2\lambda_4(\lambda_3 - \lambda_6)[\lambda_4 + 5(\lambda_3 - \lambda_6)], \tag{4.11}$$

again in agreement with Bautin, up to a scaling factor of 2π.

Before calculating a_{30} we may assume that $a_{20} = 0$, by the same argument. However, when (4.10) holds, the origin is again a center if either $\lambda_2 = 0$ or $\lambda_4 = 0$, so again there is only one choice, namely

$$\lambda_4 = -5(\lambda_3 - \lambda_6) \tag{4.12}$$

which we assume, in addition to (4.10). After a long calculation, by the formulas of § 3, the third focal value is obtained as

$$a_{30} = \tfrac{25}{64}\lambda_2(\lambda_3 - \lambda_6)^3(\lambda_3\lambda_6 - \lambda_2^2 - 2\lambda_6^2). \tag{4.13}$$

Comparing this with Bautin's formula using (4.12), we find

$$v_7 = -5(2\pi)a_{30}. \tag{4.14}$$

Thus the two formulas differ not only in scaling, but also in sign.

Evidence for a sign error in Bautin's formula came to light only recently, in the example of Shi [30] and in computer analysis by Qin and Liu [27]. The sign is important,

because it determines the stability of the weak focus of order 3, and also determines the signs of the perturbations necessary to produce the three limit cycles of Theorem 2.1.

Acknowledgments. Chengzhi Li thanks the University of Guelph for its hospitality during the course of this work. Some of the calculations in this paper were done by Nelma R. A. Moreira.

REFERENCES

[1] A. A. ANDRONOV, E. A. LEONTOVICH, I. I. GORDON, AND A. G. MAIER, *Theory of Bifurcations of Dynamical Systems in the Plane*, Israel Program for Scientific Translations, Halstead Press, John Wiley, New York, 1973.

[2] V. I. ARNOLD, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1982.

[3] R. BAMÓN, *Quadratic vector fields in the plane have a finite number of limit cycles*, Publ. Math. l'I.H.E.S., 64 (1986), pp. 111-142.

[4] N. N. BAUTIN, *On the number of limit cycles which appear with the variation of the coefficients from an equilibrium position of focus or center type*, American Mathematical Society Translation No. 100, Providence, RI, 1954. Reprinted in Stability and Dynamic Systems, American Mathematic Society Translations Series 1, Vol. 5, 1962, pp. 396-413. (Russian Original in Mat. Sb.(N.S.), 30 (1952), pp. 181-196.)

[5] YU. N. BIBIKOV, *Local Theory of Nonlinear Analytic Ordinary Differential Equations*, Lecture Notes in Mathematics 702, Springer-Verlag, Berlin, New York, 1979.

[6] G. BONIN AND Y. LEGAULT, *Comparison de la méthode des constants de Lyapunov et de la bifurcation de Hopf*, Département de Mathématics, Université de Montréal, Montréal, Quebec, Canada, 1986.

[7] L. S. CHEN AND M. S. WANG, *The relative position and number of limit cycles of a quadratic differential system*, Acta Math. Sinica, 22 (1979), pp. 751-758.

[8] S.-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, New York, 1982.

[9] W. A. COPPEL, *A survey of quadratic systems*, J. Differential Equations, 2 (1966), pp. 293-304.

[10] ———, *The limit cycle configurations of quadratic systems*, in Proc. Ninth Conference on Ordinary and Partial Differential Equations, University of Dundee.

[11] P. COULLET AND E. SPIEGEL, *Amplitude equations for systems with competing instabilities*, SIAM J. Appl. Math., 43 (1983), pp. 776-821.

[12] W. W. FARR AND R. ARIS, *Degenerate Hopf bifurcations in the* CSTR *with reactions* $A \to B \to C$, in Oscillations, Bifurcations and Chaos, F. V. Atkinson, W. F. Langford, and A. B. Mingarelli, eds., C.M.S. Conference Proceedings 8, American Mathematical Society, Providence, RI, 1987.

[13] F. GÖBBER AND K.-D. WILLAMOWSKI, *Lyapunov approach to multiple Hopf bifurcation*, J. Math. Anal. Appl., 71 (1979), pp. 333-350.

[14] M. GOLUBITSKY AND W. F. LANGFORD, *Classification and unfoldings of degenerate Hopf bifurcations*, J. Differential Equations, 41 (1981), pp. 375-415.

[15] M. GOLUBITSKY AND D. G. SCHAEFFER, *Singularities and Groups in Bifurcation Theory, Vol.* 1, Springer-Verlag, Berlin, New York, 1985.

[16] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, Berlin, New York, 1983.

[17] B. HASSARD AND Y.-H. WAN, *Bifurcation formulae derived from center manifold theory*, J. Math. Anal. Appl., 63 (1978), pp. 297-312.

[18] E. HOPF, *Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems*, Ber. Math.-Phys. Kl. Sächs Akad. Wiss. Leipzig 94 (1942), pp. 3-22. (See translation by Howard and Kopell in Marsden and McCracken, 1976).

[19] K. HUSEYIN AND P. YU, *Bifurcations associated with a simple zero eigenvalue and two pairs of pure imaginary eigenvalues*, in Oscillations, Bifurcations and Chaos, F. V. Atkinson, W. F. Langford, and A. B. Mingarelli, eds., C.M.S. Conference Proceedings 8, American Mathematical Society, Providence, RI, 1987.

[20] T. KOHDA, K. IMAMURA, AND Y. OONO, *Small-amplitude periodic solutions of the quadratic Liénard equation*, in The Theory of Dynamical Systems and Its Applications to Nonlinear Problems, H. Kawakami, ed., World Scientific, 1984, pp. 88-108.

[21] I. S. LABOURIAU, *Degenerate Hopf bifurcation and nerve impulse*, SIAM J. Math. Anal., 16 (1985), pp. 1121-1133.

[22] I. S. LABOURIAU, *Degenerate Hopf bifurcation and nerve impulse. Part* II, SIAM J. Math. Anal., this issue (1989), pp. 1-12.

[23] C. LI, *Two problems of planar quadratic systems*, Sci. Sinica Ser. A, 26 (1983), pp. 471-481. (In English.)

[24] ———, *Nonexistence of a limit cycle around a weak focus of order three for any quadratic system*, Chinese Ann. Math. Ser. B, 7 (1986), pp. 174-190. (In English.)

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

[27] Y. X. QIN AND Z. Q. LIU, *Computer deduction of formulas of differential equations*, Sci. Bull., 26 (1981), pp. 388-391.

[28] C. ROUSSEAU, *Bifurcation methods in quadratic systems*, in Oscillations, Bifurcations and Chaos, F. V. Atkinson, W. F. Langford, and A. B. Mingarelli, eds., C.M.S. Conference Proceedings 8, American Mathematical Society, Providence, RI, 1987.

[29] J. A. SANDERS AND F. VERHULST, *The Theory of Averaging*, Springer-Verlag, Berlin, New York, 1986.

[30] S. SHI, *A concrete example of the existence of four limit cycles for plane quadratic systems*, Sci. Sinica, Ser. A, 23 (1980), pp. 153-158.

[31] A. VANDERBAUWHEDE, *Local Bifurcation and Symmetry*, Research Notes in Mathematics Vol. 75, Pitman, Boston, 1982.

[32] YE YANQIAN, *Some problems in the qualitative theory of ordinary differential equations*, J. Differential Equations, 46 (1982), pp. 153-164.

[33] ———, *Recent contributions of Chinese mathematicians to the qualitative theory of polynomial differential systems*, Research Report 33, Australian National University, 1986.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

003

THE INSTABILITY OF AXISYMMETRIC SOLUTIONS IN PROBLEMS WITH SPHERICAL SYMMETRY*

PASCAL CHOSSAT† AND REINER LAUTERBACH‡

Abstract. Among all possible equilibria that may bifurcate from the trivial state for one-parameter vector fields with $O(3)$-symmetry, one generically exists, whatever the (absolutely irreducible) representation of $O(3)$ is. This state is characterized by its group of symmetry, which includes rotations about a fixed axis, and for that reason is called "axisymmetric." Recall that invariant spaces under irreducible representations of $O(3)$ have dimension $2l+1$ and are generated by spherical harmonics $Y_m^l(\theta, \phi)$, $-l \leqq m \leqq l$. If l is even, the instability of the axisymmetric solutions follows from a theorem of Ihrig and Golubitsky [*Phys. D* (1984), pp. 1-33]. If l is odd, this theorem fails because it requires a condition on the quadratic part of the Taylor expansion of the equivariant vector field, but in that case it must have a zero quadratic part. However, the linearized vector field along an axisymmetric solution is diagonal in this basis and the computation of its eigenvalues is easy once the equivariant structure of the vector field is known.

In this paper, using this idea, it is shown that two eigenvalues, namely those with eigendirections given by $m=2$ and $m=3$ in the basis of spherical harmonics, are simply related and have opposite signs whatever l.

Key words. differential equations, stability, bifurcation, spherical symmetry

AMS(MOS) subject classifications. 68F, 74D

Introduction. Bifurcations with spherical symmetry appear in a number of problems in mechanics, such as in models of the onset of convective flows inside planets and stars (Busse [1975]) and the buckling of a spherical shell (Knightly and Sather [1980]). These bifurcation problems also provide examples where the kernel V of the linearized equations is forced by symmetry to be of "high" dimension. Indeed, for each l there is a unique irreducible representation of $SO(3)$ having dimension $2l+1$, and in general the group of symmetry acts irreducibly on V. For the lower-dimensional cases ($l=1$ or 2), the stationary bifurcation problem has been solved completely, because they reduce to dimension one or two (Golubitsky and Schaeffer [1982], Chossat [1982]). For $l>2$, so far only partial results have been obtained, either by an explicit computation of the lowest-order coefficients of the bifurcation equation (for $l=3$, 4, 6, see Busse [1975] and Busse and Riahi [1982]), or by means of group-theoretic methods (Sattinger [1983], Ihrig and Golubitsky [1984]), which imply the following: to each isotropy subgroup Σ of $O(3)$ (having a one-dimensional fixed-point space in V), there corresponds a branch of bifurcated solutions whose symmetries are precisely Σ. Of course solutions in the same group orbit have isotropy subgroups conjugate to Σ. Then a study of the (conjugacy classes of) isotropy subgroups of $O(3)$ has led to a complete classification of solutions associated with such Σ's. In particular, for every $l \geqq 1$ there exists a branch of *axisymmetric* solutions, i.e, of solutions whose isotropy group contains $SO(2)$. This branch is the easiest to compute, and for a long time people dealing with the numerical solution of partial differential equations (PDE's) with spherical symmetry restricted themselves to spaces of axisymmetric functions. On the other hand, it was shown by Ihrig and Golubitsky [1984, Thm. 3.2B] that these solutions are unstable when l is even. The crucial hypothesis in their result is that the

* Received by the editors September 22, 1987; accepted for publication (in revised form) March 7, 1988.

† Département de Mathematiques, Université de Nice, Parc Valrose, F-06034 Nice, France.

‡ Institut für Mathematik, Universität Augsburg, Memminger Strasse 6, D-8900 Augsburg, Federal Republic of Germany. The work of this author was supported by Deutsche Forschungsgemeinschaft under La 525/2-1.

quadratic terms in the Taylor expansion of the bifurcation equation are not identically equal to 0. However, when l is odd the vector field must be odd and the quadratic terms are precisely null, so that the theorem does not apply in this case. On the other hand, a direct computation by Busse and Riahi [1982] of the eigenvalues of the Jacobian matrix at an axisymmetric solution in the case $l=3$ shows that at least one of these must be positive (they are real). Hence the axisymmetric solution is still unstable when $l=3$. Further calculations by hand (for $l=5\cdots$) and by computer (MacIntosh) allowed us to verify the validity of this result up to $l=15$ (the limit of our computer's capability $\cdots$). This was enough to convince us that the instability of the axisymmetric solution must be true for every odd value of l. In the present paper we present a proof of this fact. More precisely, let

$$\frac{dx}{dt}=F(x,\lambda) \tag{0.1}$$

be the equation in V (it can, for example, derive from a PDE by the center manifold reduction). We make the following hypotheses.

(H1) $O(3)$ acts absolutely irreducibly on V by its natural odd representation: $\dim V=2l+1$, l odd, and the inversion $\mathbf{x}\to-\mathbf{x}$ in $O(3)$ acts as $-\mathrm{Id}$ on V. We recall that a real representation is *absolutely irreducible* if the only linear maps that commute with it are scalar multiples of the identity. We note this representation by D^l.

(H2) $F(.,\lambda)$ is $O(3)$-equivariant, i.e., $F(gx,\lambda)=gF(x,\lambda)$ for all $g\in O(3)$, $(x,\lambda)\in V\times\mathbf{R}$.

It follows from (H1)-(H2) that: (i) 0 is an equilibrium of (0.1) for every λ; and (ii) $L_\lambda=D_xF(0,\lambda)=c(\lambda)\mathrm{Id}_V$.

We further assume the following:

(H3) $c(0)=0$ and $c'(0)\neq 0$ (we can set $c'(0)=1$).

Hypothesis (H3) implies that the "equivariant branching lemma" holds (Cicogna [1981], Vanderbauwhede [1980]): given an isotropy subgroup Σ of $O(3)$ and its fixed-point subspace $V^\Sigma=\{x\in V:\ \gamma x=x\ \forall\gamma\in\Sigma\}$, if $\dim V^\Sigma=1$ there exists a branch of solutions in V^Σ (it is straightforward algebra to check that V^Σ is invariant under equivariant F's).

THEOREM. *Let hypotheses* (H1)-(H3) *hold*; *then generically the axisymmetric equilibria of* (0.1) *bifurcating at* $\lambda=0$ *are unstable.*

Remark. We make the "generic" condition precise in § 2.2, Lemma 2. In the two next sections we (1) recall some basic group-theoretic facts and construct the axisymmetric bifurcated solutions; and (2) study the linearized equations at an axisymmetric solution and show the instability result.

1. The existence of axisymmetric solutions. In this section we indicate which solutions of an $O(3)$-equivariant bifurcation problem are axisymmetric and we compute them using the equivariant branching lemma (stated in the Introduction).

For each odd number $n=2l+1$ there exists (up to equivalence) precisely one absolutely irreducible representation of $SO(3)$ on an n-dimensional real vector space. It is equivalent to the natural representation on the space V_l of spherical harmonics of order l, which are defined in spherical coordinates by

$$Y_m^l(\theta,\phi)=P_m^l(\cos\theta)e^{mi\phi},\qquad -l\leqq m\leqq l,$$

and the P_m^l's are the associate Legendre polynomials (Miller [1972]). For the group $O(3)$ there exist two representations on V_l according to whether the inversion in $O(3)$ acts as ld or $-$ld on V. These representations are called the plus or minus representation of order l, respectively. For even l the plus representation is the natural one to occur while the minus one is natural for odd l. Recall that a spherical harmonic $h(x)$ of order l is a homogeneous polynomial on $\mathbf{R}^3$ of degree l restricted to the 2-sphere. Thus it is easily seen from the action of an element γ on a spherical harmonic $h(x)$ by

$$(\gamma h)(x) = h(\gamma x),$$

where $x \in \mathbf{S}^2$ and γ acts on x by matrix multiplication, that $(-\mathrm{ld})h(x) = (-1)^l h(x)$, and hence the natural representation of $O(3)$ is the one indicated above. We consider bifurcation problems

$$F: V_l \times \mathbf{R} \to V_l,$$

which are $O(3)$ equivariant with respect to the natural action of $O(3)$ on V_l. The existence of axisymmetric solutions follows from the following remarks and the equivariant branching lemma (Ihrig and Golubitsky [1984]): if l is *even*, all isotropy subgroups have the form $H \oplus \{-\mathrm{ld}\}$, where H is a subgroup of $SO(3)$; therefore it suffices to look at representations of $SO(3)$. The subgroup $O(2)$ of $SO(3)$ has a one-dimensional fixed point space. It is a maximal closed subgroup of $SO(3)$ and therefore it is maximal with respect to the property of having a one-dimensional fixed point space. If l is *odd* then $SO(2)$ has a one-dimensional fixed point space; however, it is not maximal with this property. The normalizer of $SO(2)$ in $SO(3)$ is $O(2)$ and it acts as minus identity on $V^{SO(2)}$. Therefore the product of an element in $O(2)\backslash SO(2)$ with $-$ld fixes $V^{SO(2)}$. The group generated by $SO(2)$ and this particular element is denoted by $O(2)^-$ and is isomorphic to $O(2)$. The spherical harmonic, which is invariant under $SO(2)$, is given by $Y_0^l(\theta, \phi)$. In either case the equivariant branching lemma guarantees a unique branch with symmetry group $O(2)$ or $O(2)^-$, respectively. We call either of these solutions the *axisymmetric* solutions.

As was pointed out in the Introduction, the axisymmetric solutions are *generically* transcritical, therefore unstable, when l is even. This result goes back to Busse [1975], whose method of proof was different. When l is odd however, the branches of axisymmetric solutions are of *pitchfork* type, because the inversion in $O(3)$ acts as $-$ld in $V^{SO(2)}$. For $l=1$ the representation is equivalent to the action of $O(3)$ in $\mathbf{R}^3$, and every element in V is axisymmetric. The problem, therefore, is similar to having a bifurcation from a simple eigenvalue, and the bifurcated solutions are stable if supercritical.

2. Instability for odd l, $l>1$.

2.1. Construction of the bifurcation equations. We briefly recall a method described in more detail in Sattinger [1979], to construct the lowest-order terms in the Taylor expansion of an equivariant bifurcation problem.

The Lie algebra $so(3)$ is generated by the matrices

$$L_1 = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{vmatrix}, \quad L_2 = \begin{vmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{vmatrix}, \quad L_3 = \begin{vmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{vmatrix}. \tag{2.1}$$

Using these operators, we define $J_+ = L_2 + iL_1$, $J_- = -L_2 + iL_1$, and $J_3 = -iL_3$. We further note

$$\gamma_m = \{(l-m)(l+m+1)\}^{1/2}. \tag{2.2}$$

We then conclude from Theorem 5.21 in Sattinger [1979] that there exists a basis $\{\xi_{-l}, \cdots, \xi_l\}$ in the complexification of V_l such that

$$J_3\xi_m = m\xi_m, \quad J_+\xi_m = \gamma_m\xi_{m+1}, \quad J_-\xi_m = \gamma_{-m}\xi_{m-1}. \tag{2.3}$$

Moreover, since V_l is real, the ξ_m's must satisfy the relation (Sattinger [1979])

$$\xi_m = (-1)^m\bar{\xi}_{-m}, \qquad -l \leqq m \leqq l. \tag{2.4}$$

These relations allow us to compute the terms in the Taylor expansion of $F(.,\lambda)$ up to any given order. First we observe that the linear term is $c(\lambda)\,\mathrm{Id}$. Second, the constant and quadratic terms are 0 because $-\mathrm{Id}$ commutes with F (l is odd). Therefore we are interested in the third-order terms. We use coordinates z_j in terms of the basis $\{\xi_j, j=-l, \cdots, l\}$. The third-order terms may be written in coordinates as

$$F_m^{(3)} = \sum_{-l\leqq r+s+t\leqq l} a_{mrst} z_r z_s z_t, \qquad -l \leqq m \leqq l. \tag{2.5}$$

The infinitesimal generators J_3, J_+, J_- act as derivations on the z_j's; hence the following relations hold for the cubic terms:

$$\begin{aligned} J_3(z_rz_sz_t) &= (r+s+t)z_rz_sz_t, \\ J_+(z_rz_sz_t) &= \gamma_r z_{r+1}z_sz_t + \gamma_s z_r z_{s+1} z_t + \gamma_t z_r z_s z_{t+1}, \\ J_-(z_rz_sz_t) &= \gamma_{-r} z_{r-1}z_sz_t + \gamma_{-s} z_r z_{s-1} z_t + \gamma_{-t} z_r z_s z_{t-1}. \end{aligned} \tag{2.6}$$

Applying J_3 to $F_m^{(3)}$ we see that $a_{mrst}=0$ if $r+s+t \neq m$. Acting with J_+ on F_m we get

$$J_+F_{m+1} = \gamma_m F_m \quad (-l \leqq m \leqq l-1), \qquad J_+F_l = 0. \tag{2.7}$$

The latter equation gives a linear system for the coefficients a_{lrst}. It is underdetermined and can be solved with respect to a certain number $n(l)$ of parameters (independent coefficients).

LEMMA 1. $n(l) = [l/3]+1$.

Proof. Let $T(l, m)$ be the set of triples $t=(i,j,k)$ of integers $-l \leqq i,j,k \leqq l$ such that $i+j+k=m$ and $i \leqq j \leqq k$. We can view the equation $J_+F_l=0$ as a linear equation from $V_{l,l}$ to $V_{l,l+1}$, where we note $V_{l,m}$ the free vector space over $T(l,m)$. We want to prove that the kernel of the associated operator A has dimension $n(l)=[l/3]+1$. Let us first define an ordering in $T(l,m)$: $(i_1,j_1,k_1) < (i_2,j_2,k_2)$ if $i_1 < i_2$ or $i_1=i_2$ and $j_1<j_2$ or $i_1=i_2$, $j_1=j_2$ and $k_1<k_2$. Then there exists an integer r such that $T(l,l+1) = \{\tau_1, \cdots, \tau_r\}$ with $\tau_i < \tau_{i+1}$, $i=1,\cdots,r-1$. We write the system to solve in the coordinates along the basis elements $\tau=(i,j,k) \in T(l,l+1)$: each coordinate equation has the form

$$\gamma_{i-1}a_{l,i-1,j,k} + \gamma_{j-1}a_{l,i,j-1,k} + \gamma_{k-1}a_{l,i,j,k-1} = 0.$$

We can now order the elements of $T(l,l)$ as follows:

(1) $\tau'_1, \cdots, \tau'_r$ where each τ' is deduced from τ by taking $\tau' = (i-1,j,k)$; (2) the remaining elements (in whatever order). Observe the following: (i) $\tau_i < \tau_j \Rightarrow \tau'_i < \tau'_j$, for all $i,j=1,\cdots,r$; and (ii) $\tau' < (i,j-1,k)$ and $\tau' < (i,j,k-1)$. It now follows that the matrix of A in this basis, truncated to the first r columns, is triangular and has only nonzero elements on its diagonal, which proves that A is surjective.

To prove the lemma it remains to show that $\dim V_{l,l} - \dim V_{l,l+1} = [l/3]+1$. Let $j^+(i,j,k) = (i,j,k+1)$ be a map from $T(l,l)$ to $T(l,l+1)$. It is not defined if $k=l$, and the image of j^+ is complementary to set $\{(i,j,j) \in T(l,l+1)\}$. There are $l+1$ elements in $T(l,l)$ with $k=l$. On the other hand, the number of triples $(i,j,j) \in T(l,l+1)$ is equal to the number of integer solutions of

$$i+2j = l+1, \quad \text{where } -l \leqq i \leqq l/3, \quad j>0.$$

If l is odd, this equals the number of even integers between $-[l/2]$ and $[(l+1)/3]$, which is $[l/2]+[[(l+1)/3]/2]+1$. From here we get

$$n(l)=l+1-[(l+1)/3]-[l/2]+[[(l+1)/3]/2]-1=[l/3]+1. \qquad \square$$

A similar argument works in the case of even l.

It follows that the size of the linear system derived from the equation $J_+F_l=0$ grows rapidly with l. Once the structure of $F_l^{(3)}$ is known, we apply J_- recursively to $F_l^{(3)}, \cdots, F_1^{(3)}$, in order to get the coefficients of $F_{l-1}^{(3)}, \cdots, F_0^{(3)}$. We remark that F_{-m} is deduced from F_m by means of (2.4). Also, it follows easily that all the coefficients a_{mrst} are real.

From (2.1) we see that the one-parameter subgroup $\exp(tJ_3)$ acts as a rotation on the two-dimensional subspace spanned by $\{\xi_k, \xi_{-k}\}$ for $k=1, \cdots, l$. Therefore it fixes the linear span of ξ_0 (i.e., $V^{SO(2)}=\operatorname{span}[\xi_0]$), and the branch of axisymmetric solutions can be computed by solving the equation in $\operatorname{span}[\xi_0]$:

$$c(\lambda)z_0+F_0^{(3)}(0, \cdots, z_0, \cdots, 0)=0. \tag{2.8}$$

The leading part of the solution is determined by the third-order terms. In the following we set

$$\beta_m=a_{mm00} \quad \text{for } m=0, \cdots, l. \tag{2.9}$$

Then we have

$$(\lambda+o(|\lambda|)z_0)+\beta_0 z_0^3+o(|z_0|^3)=0,$$

the limits being taken for $\lambda, z_0 \to 0$. After eliminating the trivial solution and applying the implicit function theorem, we can solve for

$$\lambda(z_0)=-\beta_0 z_0^2+o(|z_0|^2). \tag{2.10}$$

The next step is to express the eigenvalues of the linearization D_xF along the branch of axisymmetric solutions in terms of the β's. We consider partial derivatives

$$\frac{\partial F_m^{(3)}(0, \cdots, z_0, \cdots, 0)}{\partial z_k}, \qquad 0 \leqq k, \qquad m \leqq l. \tag{2.11}$$

If $r+k+t=m$ and $k \neq m$, at least one of the numbers r or t is nonzero and the expression in (2.11) is zero. Therefore the linearization of f along the axisymmetric solution has diagonal form (this in fact is due to the form of the Cartan decomposition of D' into irreducible representation of $O(2)^-$). If $m=k$ the expression in (2.11) is the coefficient of $z_0^2 z_m$, which is by definition β_m. Therefore the linearization near $\mathbf{z}=0$ along the axisymmetric solutions takes the form

$$-\beta_0 z_0^2 \operatorname{Id}+\operatorname{diag}(\beta_{-l}z_0^2, \cdots, 3\beta_0 z_0^2, \cdots, \beta_l z_0^2)+o(|z_0|^2). \tag{2.12}$$

The eigenvalues are

$$\sigma_m=(\beta_m-\beta_0)z_0^2+o(|z_0|^2) \quad \text{for } m=-l, \cdots, l, \quad m \neq 0, \tag{2.13}$$

and $\sigma_0=2\beta_0 z_0^2+o(|z_0|^2)$. If the quotient of two of the constants $\beta_m-\beta_0$ is negative, then the solution is unstable. Observe that two eigenvalues must be equal to 0, since the orbit of axisymmetric solutions is two-dimensional (Chossat [1982]). These eigenvalues are σ_1 and σ_{-1} since the tangent space to the orbit at z_0 is spanned by ξ_1 and ξ_{-1}. This program can be carried through numerically. We can compute $(\beta_2-\beta_0)/(\beta_3-\beta_0)$ by solving the equation $j_+F_l^{(3)}=0$, applying J_-, and reading of the β's. The symmetry of the problem seems to stabilize the numerics since, for example, for $l=15$, the linear

equation is a 85×85 matrix, where the solution depends on five parameters and still the numerical results are very precise. We give a short list of the numerical results:

l	$(\beta_2-\beta_0)/(\beta_3-\beta_0)$
3	−0.6667
5	−2.6667
7	−5.5556
9	−9.3333
11	−14.0000
13	−19.5556
15	−26.00

By the method of computation we obtain $n(l)$ numbers, which should be all equal. This numerical result gives a strong indication that all axisymmetric solutions should be unstable. The purpose of the next section will be to show that the quotient $(\beta_2-\beta_0)/(\beta_3-\beta_0)$ is always negative for $l>1$.

2.2. Computation of $\beta_2-\beta_0$ and $\beta_3-\beta_0$.

LEMMA 2. *If the coefficient of $z_{-1}^2 z_2$ in $F_0^{(3)}$ is not zero, we have*

$$\frac{\beta_2-\beta_0}{\beta_3-\beta_0}=\frac{6-l(l+1)}{9}.$$

Remark. The hypothesis in this lemma is *generic* (it holds for a wide class of problems with $O(3)$-symmetry).

Proof. We need to compute the β_k's, $k=0,2,3$, in terms of the independent coefficients appearing in $F^{(3)}$. For this we proceed as follows: (1) determine those terms in $F_0^{(3)}$ that contribute to the β_k's by successive application of the operator J_+ to $F_k^{(3)}$, $k=0,1,2$; (2) compute the equivariant relations between these terms, by means of the following expression (Miller [1972]):

$$J_-J_+F_0=l(l+1)F_0; \tag{2.14}$$

and (3) deduce from (1) and (2) the relations between the β_k's.

The part in $F_0^{(3)}$ that gives a contribution to $\beta_k z_0^2 z_k$ in $F_k^{(3)}(k=1,2,3)$ by applying relations (2.7) is

$$az_{-3}z_0z_3+bz_{-2}z_{-1}z_3+cz_{-3}z_1z_2+dz_{-2}z_0z_2+ez_{-1}^2z_2+fz_{-1}z_0z_1+gz_0^3+hz_1^2z_{-2}. \tag{2.15}$$

Of course, we have $g=\beta_0$. In addition, the following terms are needed for computing the equivariant relations between the foregoing coefficients $a,\cdots,h$ in F_0, because they are generated by applying J_+J_- to (2.15):

$$rz_{-4}z_0z_4+sz_{-4}z_1z_3+tz_{-4}z_2^2+uz_{-3}z_{-1}z_4+vz_{-2}^2z_4. \tag{2.16}$$

The terms in (2.15) are obtained in a straightforward way by applying formula (2.6) for J_- recursively, starting from $\beta_3 z_0^2 z_3$ in $F_3^{(3)}$. The terms in (2.16) are obtained by applying J_+J_- to (2.15), using (2.6). Now applying (2.14) to (2.15), we obtain a system of eight linear equations for the corresponding coefficients. Recall that the coefficients of the equivariant polynomial mappings are real, and that ξ_0 is a real vector. Therefore

the coefficients of $z_i z_j z_k$ and $z_{-i} z_{-j} z_{-k}$ $(i+j+k=0)$ must be equal. Using these remarks and formula (2.2), we finally obtain the following eight equations:

$$
\begin{aligned}
&b-c=0, \qquad e-h=0,\\
&a(\gamma_2^2+\gamma_3^2)+2b\gamma_0\gamma_2+d\gamma_2^2=-s\gamma_{-1}\gamma_{-4}-u\gamma_{-1}\gamma_{-4}-r\gamma_{-4}^2,\\
&a\gamma_0\gamma_2+b(2\gamma_1^2+\gamma_3^2)+d\gamma_0\gamma_2+2e\gamma_1\gamma_2=-u\gamma_{-3}\gamma_{-4}-2v\gamma_{-2}\gamma_{-4},\\
&a\gamma_2^2+2b\gamma_0\gamma_2+d(\gamma_1^2+\gamma_2^2)+4e\gamma_0\gamma_1+f\gamma_1^2=0, \qquad (2.17)\\
&b\gamma_1\gamma_2+d\gamma_0\gamma_1+e(\gamma_1^2+\gamma_2^2)+f\gamma_0\gamma_1=0,\\
&d\gamma_1^2+4e\gamma_0\gamma_1+f(3\gamma_0^2+\gamma_1^2)+6g\gamma_0^2=0,\\
&f\gamma_0^2+2g\gamma_0^2=0,
\end{aligned}
$$

where the γ_m's are defined in (2.2). It is clear from this that only two coefficients can be chosen independently from the others. For example, choosing d and e as these coefficients, we get the relations

$$
a=-d+2e\gamma_0/\gamma_1(1-3\gamma_0^2/\gamma_2^2), \qquad b=c=3e(\gamma_0^2-\gamma_2^2)/\gamma_1\gamma_2
$$

$$
h=d, \quad f=-d-4e\gamma_0/\gamma_1, \quad g=-f/2.
$$

It remains to apply J_+ three times to F_0. We obtain successively

$$
\beta_1=\beta_0=d/2+2e\gamma_0/\gamma_1, \quad \beta_2=d/2, \quad \beta_3=d/2-(\gamma_0/\gamma_1-3\gamma_0^3/\gamma_1\gamma_2^2)e,
$$

from which it follows that

$$
\frac{\beta_2-\beta_0}{\beta_3-\beta_0}=\frac{6-l(l+1)}{9} \qquad (2.18)
$$

provided

$$
e\neq 0. \qquad (2.19)
$$

□

Thus, when $l\geqq 3$, (2.19) is a generic condition implying that the eigenvalues σ_2 and σ_3 have opposite signs, and thus we have proved the theorem stated in the Introduction.

Acknowledgments. We began considering this problem while visiting the Institute for Mathematics and Its Applications at Minneapolis in June 1985. The theorem was proved when the authors met again at the Michigan State University at East Lansing in March 1987. We are grateful to Professor Shui-Nee Chow for giving us the opportunity to finish this work and to Professor Martin Golubitsky for helping us improve the clarity of the mathematics in the text.

REFERENCES

F. H. BUSSE [1975], *Pattern of convection in spherical shells*, J. Fluid Mech., 72, pp. 67-85.

F. H. BUSSE AND N. RIAHI [1982], *Pattern of convection in spherical shells, Part* 2, J. Fluid Mech., 123, pp. 283-301.

P. CHOSSAT [1982], *Le problème de Bénard dans une couche sphèrique*, Thèse, Université de Nice, Nice, France.

G. CICOGNA [1981], *Symmetry breakdown from bifurcation*, Lett. Nuovo Cimento, 31, pp. 600-602.

M. GOLUBITSKY AND D. SCHAEFFER [1982], *Bifurcation with O(3) symmetry including applications to the Bénard problem*, Commun. Pure Appl. Math., 35, pp. 81-111.

E. IHRIG AND M. GOLUBITSKY [1984], *Pattern selection with $O(3)$ symmetry*, Phys. D, 13, pp. 1-33.

G. H. KNIGHTLY AND D. SATHER [1980], *Buckled states of a spherical shell under uniform external pressure*, Arch. Rational Mech. Anal., 72, pp. 315-380.

W. MILLER [1972], *Symmetry Groups and Their Applications*, Academic Press, New York, London.

D. H. SATTINGER [1979], *Group Theoretic Methods in Bifurcation Theory*, Lecture Notes in Mathematics 762, Springer-Verlag, Berlin, Heidelberg, New York.

——— [1983], *Branching in the Presence of Symmetry*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

A. D. VANDERBAUWHEDE [1980], *Local bifurcation and symmetry*, Habilitation Thesis, Rijskuniversiteit Ghent, Ghent, Belgium.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

004

ON A REDUCTION PROCESS FOR NONLINEAR EQUATIONS*

A. D. JEPSON† AND A. SPENCE‡

Abstract. A method is discussed for reducing a nonlinear problem to a smaller (finite-dimensional) equivalent problem. The method is a generalization of the Lyapunov-Schmidt reduction, and provides a theoretical basis for more computationally convenient approaches. Our main results are on the equivalence of reduced problems obtained from various forms of this reduction procedure.

Key words. Lyapunov-Schmidt reduction, singularity theory, bifurcation.

AMS(MOS) subject classifications. primary 58C27; secondary 47H17

1. Introduction. Nonlinear parameter dependent problems of the form

$$F(x, \lambda, \alpha) = 0, \tag{1.1}$$

arise in the study of the equilibrium states of many nonlinear systems. Here x, the state variable, lies in some Banach space X, $\lambda \in R$ is some distinguished parameter, which is called the *bifurcation* parameter, and $\alpha \in R^p$ is a vector of *control* parameters. It is assumed that the nonlinear function F maps $X \times R \times R^p$ to Y, a Banach space, and is sufficiently smooth so that any necessary derivatives exist. In physical situations interest often centres on the critical points of (1.1), namely, the points $(x_0, \lambda_0, \alpha_0)$, say, at which $F_x^0 \equiv F_x(x_0, \lambda_0, \alpha_0)$ is singular, since it is at such points that a physical system may lose stability (see, for example, [2], [3]).

The fundamental tools for the study of solutions of (1.1) in a neighborhood of a critical point are the Lyapunov-Schmidt reduction [7], [11], [13]-[15], and the closely related alternative method [1], [5]. No matter what the form of the original equation, provided these reduction processes can be applied, a *reduced problem* is obtained of the form

$$h(\varepsilon, \lambda, \alpha) = 0, \qquad h: R^m \times R \times R^p \to R^m, \tag{1.2}$$

whose solutions are in a one-to-one correspondence with those of (1.1). Typically the dimension of the reduced problem is very small, say $m = 1$ or 2, and hence it can be readily analyzed using singularity theory. Recently this approach has led to significant advances in our understanding of nonlinear phenomena arising from a wide variety of problems (see [14] and the references cited therein). The reduction process we shall describe and discuss is a generalization of the Lyapunov-Schmidt reduction.

The standard Lyapunov-Schmidt reduction requires a precise characterization of both the null space of F_x^0 and the range of F_x^0, which we denote by $N[F_x^0]$ and $R[F_x^0]$, respectively. The particular form of the reduction process is determined by this knowledge along with choices for complementary spaces of $N[F_x^0]$ and $R[F_x^0]$. Different choices of the complementary spaces lead to different reduced problems. However, Golubitsky and Schaeffer [7] show that this nonuniqueness is unimportant in that all reduced problems derived from one original problem (through the use of

* Received by the editors April 2, 1985; accepted for publication (in revised form) February 22, 1988.

† Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4. The work of this author was supported by the Natural Sciences and Engineering Research Council of Canada and the University of Toronto.

‡ School of Mathematics, University of Bath, Claverton Down, Bath, United Kingdom, BA2 7AY. The work of this author was supported by the Natural Sciences and Engineering Research Council of Canada and the Sciences and Engineering Research Council of the United Kingdom.

the Lyapunov-Schmidt procedure) are equivalent in the sense that their solutions exhibit the same qualitative behavior. Furthermore, the solutions of the finite-dimensional reduced problem (1.2) are qualitatively similar (for a precise definition see § 3) to the solutions of the original problem (1.1). Therefore the local structure of solutions for (1.1) near a singular point $(x_0, \lambda_0, \alpha_0)$ can be studied by analyzing any reduced problem obtained through the Lyapunov-Schmidt procedure.

The motivation for this work arose in considering numerical techniques for computing the location and type of singular points for (1.1). It is hoped that these numerical techniques will form a bridge between the local results of singularity theory and the global properties often of interest in physical applications. In [9] the singularity theory of Golubitsky and Schaeffer [6] is used to derive numerically convenient defining equations for singularities arising in problems of the form (1.2) with $m = 1$. In effect it is assumed in [9] that the reduced problem (1.2) is given explicitly, as in the example calculation described there. Clearly, an important extension is to develop numerical techniques for obtaining a suitable reduction for problems of the general form (1.1). Some preliminary ideas to this end are presented in [4] and [8], and a more complete treatment is given in [10].

The fact that it is not necessary to know $N[F_x^0]$ or $R[F_x^0]$ in order to carry out a successful reduction is described in [11, § 22] and this observation leads to a generalised reduction process, described in § 2, which provides more flexibility than the standard Lyapunov-Schmidt procedure. This additional flexibility is important in the numerical computation of singular points. The typical situation in this application is that some point $(x_1, \lambda_1, \alpha_1)$, near a singular point $(x_0, \lambda_0, \alpha_0)$ of a particular type, is known. The known point need not be a singular point of (1.1), in fact, usually it is not a solution point of (1.1). However, crude approximations to $N[F_x^0]$ and $N[(F_x^0)^*]$ (and therefore $R[F_x^0]$) are available. Given this starting data, the goal of the computation is to accurately locate $(x_0, \lambda_0, \alpha_0)$. In this paper we show that crude approximations of $N[F_x^0]$ and $N[(F_x^0)^*]$ are sufficient to calculate a suitable reduced problem of the form (1.2). This result is central in the development of computationally efficient and robust numerical techniques (see [8] and [10]). Finally, we mention that the generalised reduction might also be used to advantage in analytical calculations where the null vectors of F_x^0 and $(F_x^0)^*$ are not available in simple closed form.

In § 3 we prove our main result. In particular we show that all reduced problems of the form (1.2), obtained from one original problem using the generalised procedure, are equivalent. This is an extension of the result given in [7] for the Lyapunov-Schmidt reduction. (We are grateful to a referee of this paper for supplying a shorter proof of the result for the Lyapunov-Schmidt reduction, which is presented in the Appendix.) A simple form of this result appears without proof in [8]. Independently Beyn [4] has shown that if F in (1.1) is finite-dimensional then, no matter how the generalised reduction is carried out, F is equivalent to a nonsingular linear system together with the particular reduced equation obtained. The equivalence of the reduced problems, however, is not proved there.

Finally in § 4 we consider the nonlinear differential system

$$\dot{x} + F(x, \lambda, \alpha) = 0$$

where $x \in X \times [0, \infty)$. If $(x_0, \lambda_0, \alpha_0)$ is an equilibrium solution with F_x^0 singular and all other eigenvalues have positive real part, then it is often important to ascertain the asymptotic stability of equilibrium solutions near $(x_0, \lambda_0, \alpha_0)$. If F_x^0 has zero as a simple eigenvalue Golubitsky and Schaeffer [7] show how this may be done by examination of the reduced equation derived using the Lyapunov-Schmidt reduction.

We show that the same stability information may be obtained from the generalised procedure.

2. The generalised reduction. In this section we describe the generalised reduction process as given in [11], and give necessary and sufficient conditions for its application (Theorem 2.10).

First we remark that, although in applications it is often important to keep λ a distinguished parameter, here it is convenient to put $\lambda = \alpha_0$ and write (1.1) as

$$F(x, \alpha) = 0, \qquad \alpha = (\alpha_0, \cdots, \alpha_p) \in \mathbb{R}^{p+1}. \tag{2.1}$$

Furthermore, without loss of generality, we take $(x_0, \alpha_0) = 0 \in X \times \mathbb{R}^{p+1}$, so that

$$F(0, 0) = 0. \tag{2.2}$$

Throughout this paper we assume that the Frechet derivative $L \equiv F_x(0, 0)$ satisfies

(2.3a) $L: X \to Y$ is a Fredholm operator of index zero,

with

$$N[L] = \text{span}\,\{\Phi_1, \cdots, \Phi_l\}, \qquad \dim N[L] = l, \tag{2.3b}$$

$$N[L^*] = \text{span}\,\{\Psi_1^*, \cdots, \Psi_l^*\}, \qquad \dim N[L^*] = l. \tag{2.3c}$$

Here $L^*: Y^* \to X^*$ is the adjoint operator associated with L.

As was mentioned in the Introduction, the aim of a reduction process is to obtain a small (finite) system of equations which is, in some sense, equivalent to (1.1). The first step in the generalised reduction is to choose closed subspaces X_1, $X_2 \subset X$ and Y_1, $Y_2 \subset Y$ with

$$X = X_1 \oplus X_2, \qquad \dim X_2 = m, \tag{2.4a}$$

$$Y = Y_1 \oplus Y_2, \qquad \dim Y_2 = m. \tag{2.4b}$$

Here $m \geqq l$ is taken to be finite. Let P and Q be the projections

$$P: X \to X, \quad R[P] = X_2, \quad N[P] = X_1, \tag{2.5a}$$

$$Q: Y \to Y, \quad R[Q] = Y_2, \quad N[Q] = Y_1. \tag{2.5b}$$

The fundamental assumption is that the decompositions in (2.4a), (2.4b) are such that

$$N[P] \cap N[(I - Q)L] = \{0\}. \tag{2.6}$$

Using the projections in (2.5) we rewrite (2.1) in the equivalent form

$$(I - Q)F(x_1 + x_2, \alpha) = 0 \in Y_1, \tag{2.7a}$$

$$QF(x_1 + x_2, \alpha) = 0 \in Y_2, \tag{2.7b}$$

where $x_1 \in X_i$ for $i = 1, 2$. In particular, notice that (2.7b) is a system of dimension m. We emphasize that there is no need to restrict attention to reduced problems of dimension l, and reduced problems of dimension $m \geqq l$ may be considered if deemed necessary (for example, near a singularity of higher multiplicity).

The reduction proceeds by first solving (2.7a) for x_1 in terms of x_2 and α, and, second, substituting this problem into (2.7b) to obtain

$$QF(x_1(x_2, \alpha) + x_2, \alpha) = 0. \tag{2.8}$$

The key step is the first, and the Implicit Function Theorem ensures that this is possible provided $A \equiv (I-Q)L(I-P)$ satisfies

$$A: X_1 \to Y_1 \text{ is nonsingular.} \tag{2.9}$$

Under assumption (2.6) A is indeed nonsingular. In fact, we have Theorem 2.10.

THEOREM 2.10. *Let L, P, Q, and A be as above. Then* (2.6) *is a necessary and sufficient condition for A to be nonsingular.*

The proof of Theorem 2.10 is given at the end of this section. The theorem shows that, under condition (2.6), the Implicit Function Theorem provides the existence and local uniqueness of a solution $x_1(x_2, \alpha)$ of (2.7a) for (x_2, α) near 0, and with $x_1(0,0)=0$. At this point it is convenient to introduce bases $\{v_i\}_{i=1}^m$ and $\{w_i\}_{i=1}^m$ of X_2 and Y_2, respectively, and to let $V_2: \mathbb{R}^m \to X_2$ and $W_2: Y_2 \to \mathbb{R}^m$ be defined by

$$V_2\varepsilon = \sum_{i=1}^{m} \varepsilon_i v_i, \qquad \varepsilon \equiv (\varepsilon_1, \cdots, \varepsilon_m), \tag{2.11a}$$

$$W_2\left\{\sum_{i=1}^{m} \eta_i w_i\right\} = \eta \equiv (\eta_1, \cdots, \eta_m). \tag{2.11b}$$

The final step in the reduction process is to obtain the reduced problem by writing (2.8) in terms of these bases, that is,

$$h(\varepsilon, \alpha) \equiv W_2 QF(\Omega(\varepsilon, \alpha), \alpha) = 0, \tag{2.12a}$$

where $h: \mathbb{R}^m \times \mathbb{R}^{p+1} \to \mathbb{R}^m$, and

$$\Omega(\varepsilon, \alpha) \equiv x_1(V_2\varepsilon, \alpha) + V_2\varepsilon. \tag{2.12b}$$

The Lyapunov-Schmidt reduction is a special case of the above procedure; in particular, the decompositions in (2.4) are taken to satisfy

$$X = X_1 \oplus X_2, \quad \dim X_2 = l, \quad X_2 = N[L], \tag{2.13a}$$

$$Y = Y_1 \oplus Y_2, \quad \dim Y_2 = l, \quad Y_1 = R[L]. \tag{2.13b}$$

Notice that (2.5b) and (2.13b) imply that $I-Q$ projects onto $R[L]$. Therefore (2.6) becomes simply $N[P] \cap N[L] = \{0\}$, which follows from (2.13a). That is, the decompositions used in the Lyapunov-Schmidt reduction are sufficient to guarantee (2.6) and hence that A is nonsingular. By contrast, in the generalised reduction we have more freedom in the choice of the decompositions of X and Y; in effect, the only restriction is that the decompositions produce a nonsingular A.

A second reduction technique is the alternative method [1], [5]. For the case of Fredholm operators treated here it is natural to consider only reduced problems of the form (2.12). In this situation the alternative method assumes that the decompositions (2.4a), (2.4b) are chosen such that

$$X_2 = N[(I-Q)L], \tag{2.14a}$$

$$LX_2 \subset Y_2, \tag{2.14b}$$

$$L(I-P): X_1 \to Y_1 \text{ is nonsingular.} \tag{2.14c}$$

Therefore, if L is a Fredholm operator of index zero, the alternative method is also a special case of the generalised reduction. However, it is important to note that the alternative method can also be applied to more general problems.

We end this section with the following proof.

Proof of Theorem 2.10. Define

$$\bar{A} \equiv (I-Q)L(I-P), \qquad \bar{A}: X \to Y. \tag{2.15}$$

Then $\bar{A}$ is the product of three Fredholm operators, each of which has index zero. Therefore it follows that (see [12]),

(2.16) $\bar{A}$ is Fredholm with index zero.

Assume that (2.6) is satisfied, that is,

(2.17) $$R[I-P]\cap N[(I-Q)L]=\{0\}.$$

By (2.15) we have

(2.18) $$N[\bar{A}]=N[I-P]\oplus\{R[I-P]\cap N[(I-Q)L]\}.$$

Therefore, by (2.5a), (2.17), and (2.18),

(2.19a) $$N[\bar{A}]=X_2, \qquad \dim X_2=m.$$

Furthermore, by (2.15), we have $R[\bar{A}]\subset R[I-Q]=X_1$. But by (2.16) and (2.19a) we have codim $R[\bar{A}]=m$. It follows from (2.4b) and (2.5b) that codim $R[I-Q]=m$, and hence

(2.19b) $$R[\bar{A}]=R[I-Q]=Y_1, \qquad \operatorname{codim} R[\bar{A}]=m.$$

It now follows that $A=\bar{A}|_{X_1}$ is one-to-one and onto Y_1. By the Closed Graph Theorem we conclude that A is nonsingular, and therefore (2.6) is sufficient.

Conversely, suppose that $\phi\in N[P]\cap N[(I-Q)L]$, $\phi\neq 0$. Then $\phi=(I-P)\phi$ and $A\phi=\bar{A}\phi=(I-Q)L\phi=0$. That is $\phi\in N[A]$, and hence (2.6) is necessary. □

3. Equivalence of the reduced problems. The reduction processes discussed in § 2 all have the same goal, namely, the construction of a finite-dimensional reduced problem from (1.1). The type and unfolding behavior of a singular point of F is to be determined by studying the singular behavior of the reduced problem. However, the reduction process is not unique; in particular there are infinitely many ways to choose the decompositions in (2.4) such that (2.6) is satisfied. The usefulness of the reduction process therefore lies, in part, in the fact that the singular behavior of the reduced problem depends only on the function F and not on the details of the reduction. We make this precise through the use of the following notion of equivalence.

DEFINITION 3.1. Suppose $h(\varepsilon,\alpha)$, $g(\varepsilon,\alpha):\mathbb{R}^m\times\mathbb{R}^{p+1}\to\mathbb{R}^m$ are two smooth functions such that $h(0,0)=g(0,0)=0$. Then h and g are said to be equivalent, which we denote by $h\sim g$, if there exist smooth (C^∞) functions $T(\varepsilon,\alpha)$ and $E(\varepsilon,\alpha)$ such that

(3.2a) $T(0,0)$ is nonsingular,

(3.2b) $E(0,0)=0$, $E_\varepsilon^0\equiv E_\varepsilon(0,0)$ is nonsingular,

and, for (ε,α) near $(0,0)$,

(3.3) $$h(\varepsilon,\alpha)=T(\varepsilon,\alpha)g(E(\varepsilon,\alpha),\alpha).$$

For the case in which F is C^∞ in a neighbourhood of zero then (3.3) essentially states that h and g are *contact equivalent* in the sense of the singularity theory of Golubitsky and Schaeffer [6]. To be precise, (3.3) states that h is contact equivalent to $R_1 g(R_2\varepsilon,\alpha)$ for constant matrices R_1 and R_2 chosen so that $T(0,0)R_1^{-1}$ and $E_\varepsilon^0 R_2$ are positive definite. These matrices are important when the dynamical stability of various steady states are investigated in § 4; however, we can ignore them for the moment. As discussed in [6], this notion of contact equivalence is an appropriate mathematical formulation of the statement that solutions of (3.2a) and (3.2b) show the same *qualitative behavior* near $(\varepsilon,\alpha)=(0,0)$. Our main result is Theorem 3.4.

THEOREM 3.4 (Equivalence Theorem). *Let*

$$h(\varepsilon, \alpha) = 0, \qquad h : \mathbb{R}^m \times \mathbb{R}^{p+1} \to \mathbb{R}^m, \tag{3.4a}$$

$$\hat{h}(\hat{\varepsilon}, \alpha) = 0, \qquad \hat{h} : \mathbb{R}^{\hat{m}} \times \mathbb{R}^{p+1} \to \mathbb{R}^{\hat{m}} \tag{3.4b}$$

be two reduced problems obtained from (1.1) *by the generalised reduction procedure with* (2.6) *satisfied in each case. Suppose* $\hat{m} \leqq m$ *and let* $k : \mathbb{R}^m \times \mathbb{R}^{p+1} \to \mathbb{R}^m$ *be the trivial extension of* $\hat{h}$ *given by*

$$k(\varepsilon, \alpha) = (\hat{h}(\varepsilon_1, \cdots, \varepsilon_{\hat{m}}, \alpha), \quad e_{\hat{m}+1}, \cdots, \varepsilon_m). \tag{3.5}$$

Then

$$h(\varepsilon, \alpha) \sim k(\varepsilon, \alpha). \tag{3.6}$$

Before proving the theorem we first note that a similar result has been proved in [7] for the case of the Lyapunov-Schmidt reduction applied to problems with one-dimensional null spaces (i.e., (2.3) is satisfied with $l = 1$). Our proof of the more general result is guided by the proof presented there.

Proof of Theorem 3.4. It is convenient to first consider the case in which the dimensions of the reduced systems are the same, that is, $\hat{m} = m$. In this case we need to show $h \sim \hat{h}$. It is obvious that if h and $\hat{h}$ are obtained through the use of the same P and Q but with different choices of V_2 and W_2 in (2.9), then $h \sim \hat{h}$. (Use $T_{(\varepsilon,\alpha)} = T_0$ and $E(\varepsilon, \alpha) = E_0 \varepsilon$ where T_0 and E_0 are constant $m \times m$ matrices.) An h obtained from a particular choice of P and Q is denoted by h_{PQ}. Our first major task, then, is to show

$$h_{PQ} \sim h_{\hat{P}\hat{Q}} \tag{3.7}$$

whenever P, Q and $\hat{P}$, $\hat{Q}$ satisfy (2.4), (2.5), and (2.6) with the same value of m.

In order to prove (3.7) we first show that a third projection $\bar{P}$ exists such that $h_{\bar{P}Q}$ and $h_{\bar{P}\hat{Q}}$ can be constructed (see Lemma 3.8 below). Second, we show that $h_{PQ} \sim h_{\bar{P}Q}$ and $h_{\hat{P}\hat{Q}} \sim h_{\bar{P}\hat{Q}}$ (see Lemma 3.9). Finally (3.7) follows by showing that $h_{\bar{P}Q} \sim h_{\bar{P}\hat{Q}}$ (see Lemma 3.19).

LEMMA 3.8. *Assume that* P, Q *and* $\hat{P}$, $\hat{Q}$ *are two pairs of projections with each projection having a range of dimension* m, *and with each pair satisfying* (2.5) *and* (2.6). *Then there is a continuous projection* $\bar{P} : X \to X$ *such that*

$$\dim R[\bar{P}] = m, \tag{3.8a}$$

$$N[\bar{P}] \cap N[(I - Q)L] = \{0\}, \tag{3.8b}$$

$$N[\bar{P}] \cap N[(I - \hat{Q})L] = \{0\}. \tag{3.8c}$$

Proof. Let $X_3 = N[(I - Q)L]$, and similarly define $\hat{X}_3$ using $\hat{Q}$. In order to calculate $\dim X_3$, notice that $(I - Q)L$ is a product of two Fredholm operators with index zero, and hence $(I - Q)L$ is Fredholm with index zero. Therefore

$$\dim X_3 = \operatorname{codim} R[(I - Q)L] \geqq \operatorname{codim} R[I - Q] = m.$$

However, by (2.6), we have

$$\dim X_3 \leqq \operatorname{codim} X_1 = m.$$

Therefore $\dim X_3 = m$, and a similar argument with $\hat{Q}$ shows that $\dim \hat{X}_3 = m$.

Now we choose $\bar{X}_1$ such that it is the simultaneous complement of X_3 and $\hat{X}_3$. To be precise, $\bar{X}_1$ is a closed subspace of X such that

$$X = \bar{X}_1 \oplus X_3 = \bar{X}_1 \oplus \hat{X}_3.$$

(To prove that such an $\bar{X}_1$ exists, consider $\bar{X}_1$ to be the direct sum of a finite-dimensional space and a complement of $X_3 \oplus \hat{X}_3$. This reduces the problem to the finite-dimensional space $X_3 \oplus \hat{X}_3$. We omit the remaining details.) Finally, let $\bar{X}_2$ be a complement of $\bar{X}_1$, say X_3, and define $\bar{P}$ to be the projection onto $\bar{X}_2$ along $\bar{X}_1$. Conditions (3.8a)-(3.8c) easily follow.

We remark that if $l = m$ then $\bar{P}$ can be taken to be the projection used in the standard Lyapunov-Schmidt procedure. □

The generalised reduction procedure can be used to construct $h_{\bar{P}Q}$ and $h_{\bar{P}\hat{Q}}$, since Lemma 3.8 ensures that conditions (2.4) and (2.6) are satisfied for the appropriate pairs of projections. We now have Lemma 3.9.

LEMMA 3.9. *In the notation used above,*

$$h_{PQ}(\varepsilon, \alpha) \sim h_{\bar{P}Q}(\varepsilon, \alpha), \tag{3.9a}$$

$$h_{\bar{P}\hat{Q}}(\varepsilon, \alpha) \sim h_{\bar{P}\hat{Q}}(\varepsilon, \alpha). \tag{3.9b}$$

Proof. We begin by considering (3.9a). In order to obtain h_{PQ} and $h_{\bar{P}Q}$ we must solve

$$(I - Q)F(x_1 + x_2, \alpha) = 0, \tag{3.10a}$$

$$(I - Q)F(\bar{x}_1 + \bar{x}_2, \alpha) = 0, \tag{3.10b}$$

for $x_1 \in X_1$ and $\bar{x}_1 \in \bar{X}_1$, respectively. As was mentioned in § 2, locally unique smooth solutions $x_1(x_2, \alpha)$ and $\bar{x}_1(\bar{x}_2, \alpha)$ can be obtained for (x_2, α) and $(\bar{x}_2, \alpha)$ near $(0, 0)$. With these solutions, define

$$u(x_2, \alpha) = (I - \bar{P})\{x_1(x_2, \alpha) + x_2\}, \tag{3.11a}$$

$$\mu(x_2, \alpha) = \bar{P}\{x_1(x_2, \alpha) + x_2\}. \tag{3.11b}$$

The local uniqueness of $\bar{x}_1(\bar{x}_2, \alpha)$ as a solution of (3.10b) can now be used to show that

$$u(x_2, \alpha) = \bar{x}_1(\bar{x}_2, \alpha) \quad \text{for } \mu(x_2, \alpha) = \bar{x}_2. \tag{3.12}$$

Therefore, by (3.11) and (3.12) we have

$$\begin{aligned} QF(x_1(x_2, \alpha) + x_2, \alpha) &= QF(u(x_2, \alpha) + \mu(x_2, \alpha), \alpha) \\ &= QF(\bar{x}_1(\mu(x_2, \alpha), \alpha) + \mu(x_2, \alpha), \alpha). \end{aligned}$$

From this and (2.12) it follows that

$$h_{PQ}(\varepsilon, \alpha) = h_{\bar{P}Q}(E(\varepsilon, \alpha), \alpha), \tag{3.13a}$$

with

$$E(\varepsilon, \alpha) = \bar{V}_2^{-1}\mu(V_2\varepsilon, \alpha). \tag{3.13b}$$

Here $\bar{V}_2 : \mathbb{R}^m \to \bar{X}_2$ is the coordinate function (see (2.11a)) used in constructing $h_{\bar{P}Q}$.

We are left with showing that E_ε^0 is nonsingular. To do this, notice that (3.13b) gives

$$\bar{V}_2 E_\varepsilon^0 = \mu_{x_2}^0 V_2. \tag{3.14a}$$

Furthermore, by (3.11b) we find

$$\mu_{x_2}^0 = \bar{P}\left\{\frac{\partial x_1}{\partial x_2}(0, 0) + I\right\}. \tag{3.14b}$$

Let $\phi \in N[E_\varepsilon^0]$; then by (3.14),

$$0 = \bar{V}_2 E_\varepsilon^0 \phi = \bar{P}\left\{\frac{\partial x_1}{\partial x_2}(0, 0) + I\right\} V_2 \phi. \tag{3.15}$$

By differentiating (3.10a) with respect to x_2 we find

$$x(\phi) \equiv \left\{ \frac{\partial x_1}{\partial x_1}(0,0) + I \right\} V_2 \phi \in N[(I-Q)L]. \tag{3.16}$$

Together, (3.8b), (3.15), and (3.16) imply

$$x(\phi) \in N[\bar{P}] \cap N[(I-Q)L] = \{0\}.$$

That is, $x(\phi) = 0$ and therefore, by applying P to (3.16), we obtain $V_2\phi = 0$. Since V_2 is nonsingular we finally have $\phi = 0$, and therefore E_ε^0 is nonsingular.

The proof of (3.9b) is similar, with (3.8c) being used in place of (3.8b). □

We are left with proving that $h_{\bar{P}Q} \sim h_{\bar{P}\hat{Q}}$. To do this we will use the following proposition, which is the analogue of Proposition I, 4.2 in [7] for vector-valued functions.

PROPOSITION 3.17. *Let $g(\varepsilon, \alpha)$ and $\hat{g}(\varepsilon, \alpha)$ be two smooth functions, g, $\hat{g}: \mathbb{R}^m \times \mathbb{R}^{p+1} \to \mathbb{R}^m$. Suppose*

$$\operatorname{rank} \left\{ \frac{\partial g(0,0)}{\partial(\varepsilon, \alpha)} \right\} = \operatorname{rank} \left\{ \frac{\partial \hat{g}(0,0)}{\partial(\varepsilon, \alpha)} \right\} = m, \tag{3.17a}$$

and, for (ε, α) in a neighbourhood of zero,

$$g(\varepsilon, \alpha) = 0 \text{ implies } \hat{g}(\varepsilon, \alpha) = 0. \tag{3.17b}$$

Then there exists a nonsingular $m \times m$ matrix $T(\varepsilon, \alpha)$ such that

$$\hat{g}(\varepsilon, \alpha) = T(\varepsilon, \alpha) g(\varepsilon, \alpha), \tag{3.18}$$

for (ε, α) in a neighbourhood of zero. Furthermore, $T(\varepsilon, \alpha)$ inherits the smoothness of g and $\hat{g}$.

A proof of Proposition 3.17 can be obtained by a straightforward generalisation of the proof of Proposition I, 4.2 given in [7]. We omit the details. This proposition is used in the proof of Lemma 3.19.

LEMMA 3.19. *Let $h_{\bar{P}Q}$ and $h_{\bar{P}\hat{Q}}$ be as above. Then $h_{\bar{P}Q} \sim h_{\bar{P}\hat{Q}}$.*

Proof. If $h_{\bar{P}Q}(\varepsilon, \alpha) = 0$ then $F(\bar{x}_1(\bar{V}_2\varepsilon, \alpha) + \bar{V}_2\varepsilon, \alpha) = 0$, and so $h_{\bar{P}\hat{Q}}(\varepsilon, \alpha) \equiv \hat{W}_2\hat{Q}F(\bar{x}_1(\bar{V}_2\varepsilon, \alpha) + \bar{V}_2\varepsilon, \alpha) = 0$. Therefore (3.17b) is satisfied for $g \equiv h_{\bar{P}Q}$ and $\hat{g} \equiv h_{\bar{P}\hat{Q}}$. In order to apply Proposition 3.17 we must arrange for (3.17a) to be satisfied.

Specifically, we unfold F by writing

$$G(x, \alpha, \beta) \equiv F(x, \alpha) + B\beta = 0, \tag{3.20}$$

where $\beta \in \mathbb{R}^m$ and $B: \mathbb{R}^m \to Y$ is a bounded linear map. Let $g(\varepsilon, \alpha, \beta)$ and $\hat{g}(\varepsilon, \alpha, \beta)$ be the reduced functions obtained by using $\bar{P}$, Q and $\bar{P}$, $\hat{Q}$, respectively, on problem (3.20). In particular, we have

$$g(\varepsilon, \alpha, 0) = h_{\bar{P}Q}(\varepsilon, \alpha), \tag{3.21a}$$

$$\hat{g}(\varepsilon, \alpha, 0) = h_{\bar{P}\hat{Q}}(\varepsilon, \alpha), \tag{3.21b}$$

for (ε, α) near 0. Furthermore, from (3.20) it follows that

$$g_\beta^0 = KB, \; \hat{g}_\beta^0 = \hat{K}B, \tag{3.22}$$

where K, $\hat{K}: Y \to \mathbb{R}^m$ are given by

$$K \equiv [-W_2 Q F_x^0 A_{\bar{P}Q}^{-1}(I-Q) + W_2 Q], \qquad A_{\bar{P}Q} \equiv (I-Q)F_x^0(I-\bar{P}), \tag{3.23a}$$

$$\hat{K} \equiv [-\hat{W}_2 \hat{Q} F_x^0 A_{\bar{P}\hat{Q}}^{-1}(I-\hat{Q}) + \hat{W}_2 \hat{Q}], \qquad A_{\bar{P}\hat{Q}} \equiv (I-\hat{Q})F_x^0(I-\bar{P}). \tag{3.23b}$$

We need to choose B such that g_β^0 and $\hat{g}_\beta^0$ are nonsingular. This can be done as follows. By (2.11b) and (3.23) we see that

$$K Y_2 = W_2 Y_2 = \mathbb{R}^m, \qquad \hat{K}\hat{Y}_2 = \mathbb{R}^m, \tag{3.24}$$

and

$$\operatorname{codim} N[K] = m = \operatorname{codim} N[\hat{K}].$$

It is easily proved that there exists a simultaneous complement, $\bar{Y}_2$, such that

$$\begin{aligned} &N[K] \oplus \bar{Y}_2 = Y = N[\hat{K}] \oplus \bar{Y}_2, \\ &\dim \bar{Y}_2 = m, \end{aligned} \tag{3.25}$$

and hence we may define B such that $R[B] = \bar{Y}_2$. It now follows from (3.22), (3.24), and (3.25) that

$$\operatorname{rank}[g_\beta^0] = \operatorname{rank}[\hat{g}_\beta^0] = m, \tag{3.26}$$

as desired.

Now Proposition 31.7 implies that

$$\hat{g}(\varepsilon, \alpha, \beta) = T(\varepsilon, \alpha, \beta) g(\varepsilon, \alpha, \beta) \tag{3.27}$$

for $(\varepsilon, \alpha, \beta)$ near zero. By setting $\beta = 0$ it follows from (3.21) that

$$h_{\bar{P}\hat{Q}}(\varepsilon, \alpha) = T(\varepsilon, \alpha, 0) h_{\bar{P}Q}(\varepsilon, \alpha)$$

for (ε, α) near zero. Furthermore, by differentiating (3.27), it follows that $\hat{g}_\beta^0 = T^0 g_\beta^0$, with (3.26) ensuring that T^0 is nonsingular. □

Lemmas 3.8, 3.9, and 3.19 complete the proof that $h_{PQ} \sim h_{\hat{P}\hat{Q}}$ when $\hat{m} = m$. That is, for a given $m \geqq l$ the particular choice of P and Q does not effect the qualitative behavior of the reduced equation near $(\varepsilon, \alpha) = 0$.

The next major step is to show the equivalence of the reduced problems (3.4a), (3.4b) obtained with $\hat{m} < m$. In fact, it is sufficient to consider only the case $l = \hat{m} < m$, since all other cases can be derived from this case in a straightforward manner.

In view of the above results for $\hat{m} = m$, we are free to choose convenient projections P, Q and $\hat{P}$, $\hat{Q}$ to define h and $\hat{h}$. In particular, we take the splittings

$$X = X_1 \oplus X_2 \oplus X_3 \tag{3.28a}$$

$$Y = Y_1 \oplus Y_2 \oplus Y_3 \tag{3.28b}$$

where

$$N[F_x^0] = X_3, \qquad \dim X_2 = m - l, \tag{3.28c}$$

$$R[F_x^0] = Y_1 \oplus Y_2, \qquad \dim Y_2 = m - l, \tag{3.28d}$$

$$Y_2 = F_x^0 X_2. \tag{3.28e}$$

We choose $\hat{P}$ and $\hat{Q}$ such that

$$\hat{P}: X \to X_3, \qquad N[\hat{P}] = X_1 \oplus X_2, \tag{3.29a}$$

$$\hat{Q}: Y \to Y_3, \qquad N[\hat{Q}] = R[F_x^0] = Y_1 \oplus Y_2 \tag{3.29b}$$

(which corresponds to the standard Lyapunov-Schmidt reduction). Similarly, we choose P and Q such that

$$P: X \to X_2 \oplus X_3, \qquad N[P] = X_1, \tag{3.30a}$$

$$Q: Y \to Y_2 \oplus Y_3, \qquad N[Q] = Y_1 \tag{3.30b}$$

(which corresponds to the alternative method [1], [5]). It is straightforward to check that both pairs of projections satisfy (2.6), and therefore h_{PQ} and $h_{\hat{P}\hat{Q}}$ exist. The proof of Theorem 3.4 is now completed by the use of Lemma 3.31.

LEMMA 3.31. *In the above notation*

$$h_{PQ}(\varepsilon_1, \cdots, \varepsilon_m, \alpha) \sim (h_{\hat{P}\hat{Q}}(\varepsilon_1, \cdots, \varepsilon_l, \alpha), \varepsilon_{l+1}, \cdots, \varepsilon_m). \tag{3.31}$$

Proof. The reduced equation $h_{PQ} = 0$ can be written as (see 2.12)

$$h_{PQ}(\sigma, \delta, \alpha) \equiv \begin{pmatrix} W_2(I-Q)Q \\ W_2\hat{Q}Q \end{pmatrix} F(\hat{x}(\sigma, \delta, \alpha), \alpha) = 0 \tag{3.32}$$

where $\varepsilon = (\sigma, \delta)$, $\sigma \in \mathbb{R}^{m-l}$, $\delta \in \mathbb{R}^l$, and

$$W_2 : Y_2 \to \mathbb{R}^{m-l}, \qquad W_3 : Y_3 \to \mathbb{R}^l, \tag{3.33a}$$

$$V_2 : \mathbb{R}^{m-l} \to X, \qquad V_3 : \mathbb{R}^l \to X_3, \tag{3.33b}$$

are nonsingular linear mappings. Furthermore, $\hat{x}(\sigma, \delta, \alpha)$ satisfies

$$(I-Q)F(\hat{x}(\sigma, \delta, \alpha), \alpha) = 0, \tag{3.34a}$$

$$\hat{x}(\sigma, \delta, \alpha) = x_1(\sigma, \delta, \alpha) + V_2\sigma + V_3\delta, \qquad x_1 \in X_1, \tag{3.34b}$$

for (σ, δ, α) near zero.

We claim that, with the above definitions,

$$h_{PQ}(\sigma, \delta, \alpha) \sim \begin{pmatrix} \sigma \\ h_0(\delta, \alpha) \end{pmatrix} \tag{3.35}$$

for some smooth function $h_0 : \mathbb{R}^l \times \mathbb{R}^{p+1} \to \mathbb{R}^l$. Indeed, upon differentiating (3.32) with respect to σ, we find

$$H^0 \equiv \left(\frac{\partial}{\partial \sigma} h_{PQ}\right)^0 = \begin{pmatrix} W_2(Q-\hat{Q}) \\ W_2\hat{Q} \end{pmatrix} F_x^0 V_2. \tag{3.36}$$

Here we have used $\hat{Q}Q = \hat{Q}$ (see (3.29b), (3.30b)), and $\partial x_1/\partial \sigma\ (0, 0, 0) = 0$ (see (3.28), (3.30), (3.34)). It is easy to verify that the top block of H^0, namely $W_2(Q-\hat{Q})F_x^0 V_2$, is a nonsingular $(m-l) \times (m-l)$ matrix. A standard argument now ensures the existence of a smooth coordinate transformation $(\sigma, \delta, \alpha) \to (D(\sigma, \delta, \alpha), \delta, \alpha)$ such that

$$D^0 = 0, \qquad D_\delta^0 = 0, \tag{3.37a}$$

$$D_\sigma^0 \quad \text{is nonsingular} \tag{3.37b}$$

and

$$h_{PQ}(D(\sigma, \delta, \alpha), \delta, \alpha) = \begin{pmatrix} \sigma \\ W_2\hat{Q}F(\hat{x}(D(\sigma, \delta, \alpha), \delta, \alpha), \alpha) \end{pmatrix} \tag{3.38}$$

for (σ, δ, α) near zero. Finally it follows from (3.38) that there exists a nonsingular matrix $T_1(\sigma, \delta, \alpha)$ such that

$$T_1(\sigma, \delta, \alpha) h_{PQ}(D(\sigma, \delta, \alpha), \delta, \alpha) = \begin{pmatrix} \sigma \\ h_0(\delta, \alpha) \end{pmatrix}$$

where

$$h_0(\delta, \alpha) = W_2\hat{Q}F(\hat{x}(D(0, \delta, \alpha), \delta, \alpha), \alpha). \tag{3.39}$$

That is, we have verified (3.35) for h_0 as in (3.39).

We are only left with relating h_0 to $h_{\hat{P}\hat{Q}}$. Note that (3.34) and (3.38) imply that $\hat{x}(D(0, \delta, \alpha), \delta, \alpha)$ satisfies

$$(I-Q)F(\hat{x}, \alpha)=0 \quad \text{and} \quad (I-\hat{Q})QF(\hat{x}, \alpha)=(Q-\hat{Q})F(\hat{x}, \alpha)=0.$$

Hence

$$(I-\hat{Q})F(\hat{x}, \alpha)=0. \tag{3.40a}$$

In addition, we find from (3.34b) that

$$\hat{P}\hat{x}=V_3\delta. \tag{3.40b}$$

These are precisely the equations that need to be solved to obtain $h_{\hat{P}\hat{Q}}$, and therefore it follows from (3.39) and (3.40) that

$$h_0(\delta, \alpha)=h_{\hat{P}\hat{Q}}(\delta, \alpha) \tag{3.41}$$

(where V_3 and W_3 are used as the coordinate mappings in $h_{\hat{P}\hat{Q}}$). The equivalence (3.31) is a consequence of (3.35) and (3.41). This completes the proof of Lemma 3.31, and hence the proof of the Equivalence Theorem. □

4. Linearized stability and the generalised reduction. Consider now the case that (1.1) arises in the study of the differential system

$$\dot{x}+F(x, \alpha)=0, \qquad \alpha \in \mathbb{R}^{p+1} \tag{4.1}$$

where $x=x(t)\in X$, $t>0$ and we again write $\lambda=\alpha_0$. It is convenient to take $X=Y$ a Hilbert space. Assume $(0, 0)$ is an equilibrium solution. It is well known that the asymptotic stability of $(0, 0)$ depends on the spectrum of $L=F_x(0, 0)$ and that an equilibrium solution that is a singular point of L, with all other eigenvalues of L having positive real part, lies on a stability boundary. Golubitsky and Schaeffer [7, §4, Chap. I] consider such a case with L having zero as a simple eigenvalue and discuss the stability of an equilibrium solution (x, α) near $(0, 0)$. In particular, suppose Φ_0 and Ψ_0^* are the null vectors of F_x^0 and $(F_x^0)^*$, respectively, with

$$\Psi_0^*\Phi_0>0. \tag{4.2}$$

(Here and in the sequel we denote the inner product $\langle\Psi_0, \Phi_0\rangle$ by $\Psi_0^*\Phi_0$.) Furthermore, suppose $h(\varepsilon, \alpha)$ is obtained from the Lyapunov-Schmidt reduction (use $W_2=\Psi_0^*$, $V_2=\Phi_0$ in § 2); then it is shown in [7] that the solution $(x, \alpha)=(\Omega(\varepsilon, \alpha), \alpha)$ of (1.1) is stable if $h_\varepsilon(\varepsilon, \alpha)>0$ and unstable if $h_\varepsilon(\varepsilon, \alpha)<0$. The main result of this section is to show how the analogous stability results can be obtained from a reduced function computed through the use of the generalised reduction.

In the remainder of this section we assume that Ψ_0^* and Φ_0 are as above with $\Psi_0^*\Phi_0\neq 0$, but we do not assume (4.2) is satisfied. Then zero is a simple eigenvalue of F_x^0 and it follows (from the Implicit Function Theorem) that there are smooth functions $\Phi(\varepsilon, \alpha)$, $\Psi(\varepsilon, \alpha)$, and $\mu(\varepsilon, \alpha)$ such that

$$F_x(\Omega(\varepsilon, \alpha), \alpha)\Phi(\varepsilon, \alpha)=\mu(\varepsilon, \alpha)\Phi(\varepsilon, \alpha), \tag{4.3a}$$

$$\Phi^*\Phi=1, \tag{4.3b}$$

$$\Psi^*F_x(\Omega(\varepsilon, \alpha), \alpha)=\mu(\varepsilon, \alpha)\Psi^*(\varepsilon, \alpha), \tag{4.3c}$$

$$\Psi^*\Psi=1, \tag{4.3d}$$

with $\mu(0, 0)=\mu^0=0$, $\Phi(0, 0)=\Phi_0$, $\Psi(0, 0)=\Psi_0$.

As discussed above, in order to determine the (linear) stability of the steady state $(x, \alpha)=(\Omega(\varepsilon, \alpha), \alpha)$ we need only know the sign of $\mu(\varepsilon, \alpha)$, which, in turn, is to be determined from the reduced function.

Suppose $h: \mathbb{R}^m + \mathbb{R}^{p+1} \to \mathbb{R}^m$ is a reduced function obtained from (1.1) through the use of the generalised reduction. Then the connection between the null spaces of h_ε and F_x is given in Lemma 4.4.

LEMMA 4.4. *For* $\phi, \psi \in \mathbb{R}^m$ *define*

$$\Phi(\varepsilon, \alpha, \phi) \equiv \Omega_\varepsilon(\varepsilon, \alpha)\phi = (I - P)\Omega_\varepsilon(\varepsilon, \alpha)\phi + V_2\phi, \tag{4.4a}$$

$$\Psi^*(\varepsilon, \alpha, \psi) \equiv \psi^* W_2 Q[I - F_x(\Omega(\varepsilon, \alpha), \alpha)(I - P)A^{-1}(\varepsilon, \alpha)(I - Q)]. \tag{4.4b}$$

Here $A(\varepsilon, \alpha): X_1 \to Y_1$ *is defined by* (*cf.* (2.9))

$$A(\varepsilon, \alpha) \equiv (I - Q)F_x(\Omega(\varepsilon, \alpha), \alpha)(I - P). \tag{4.4c}$$

Then, for (ε, α) *near* $(0, 0)$,

$$N[F_x(\Omega(\varepsilon, \alpha), \alpha)] \subset \Phi(\varepsilon, \alpha, \mathbb{R}^m) \equiv \{\Phi(\varepsilon, \alpha, \phi) | \phi \in \mathbb{R}^m\}, \tag{4.5a}$$

$$N[(F_x(\Omega(\varepsilon, \alpha), \alpha))^*] \subset \Psi^*(\varepsilon, \alpha, \mathbb{R}^m). \tag{4.5b}$$

Furthermore,

$$h_\varepsilon(\varepsilon, \alpha)\phi = 0, \tag{4.6a}$$

$$\psi^* h_\varepsilon(\varepsilon, \alpha) = 0, \tag{4.6b}$$

if and only if

$$F_x(\Omega(\varepsilon, \alpha), \alpha)\Phi(\varepsilon, \alpha, \phi) = 0, \tag{4.7a}$$

$$\Psi^*(\varepsilon, \alpha, \psi)F_x(\Omega(\varepsilon, \alpha), \alpha) = 0. \tag{4.7b}$$

Proof. First note that $A(0, 0)$ is precisely the A used in (2.9), which is assumed to be nonsingular, and therefore $A^{-1}(\varepsilon, \alpha)$ exists for (ε, α) near $(0, 0)$. Also, recall from § 2 that $\Omega(\varepsilon, \alpha)$ is defined to be the solution of

$$(I - Q)F(\Omega(\varepsilon, \alpha), \alpha) = 0, \tag{4.8a}$$

$$P\Omega(\varepsilon, \alpha) = V_2\varepsilon, \tag{4.8b}$$

and the reduced function is

$$h(\varepsilon, \alpha) = W_2 QF(\Omega(\varepsilon, \alpha), \alpha). \tag{4.8c}$$

By differentiating (4.7) we obtain

$$(I - Q)F_x(\Omega(\varepsilon, \alpha), \alpha)\Omega_\varepsilon(\varepsilon, \alpha) = 0, \tag{4.9a}$$

$$P\Omega_\varepsilon(\varepsilon, \alpha) = V_2, \tag{4.9b}$$

$$h_\varepsilon(\varepsilon, \alpha) = W_2 QF_x(\Omega(\varepsilon, \alpha), \alpha)\Omega_\varepsilon(\varepsilon, \alpha). \tag{4.9c}$$

Equations (4.9a) and (4.9b) can be solved to give

$$\Omega_\varepsilon(\varepsilon, \alpha) = [I - (I - P)A^{-1}(\varepsilon, \alpha)(I - Q)]V_2, \tag{4.10}$$

which we make use of below.

Now suppose (4.6) is satisfied. Then (4.7a) clearly follows from (4.4a), (4.6a), (4.9a), (4.9c), and the nonsingularity of W_2. In order to show that (4.7b) follows from (4.6b), we rewrite (4.7b) as

$$\Psi^* F_x = \Psi^* F_x(I - P) + \Psi^* F_x P = 0$$

and consider the two components in the sum separately. By (4.4b) and (4.4c) we have

$$\Psi^* F_x(I - P) = \psi^* W_2 Q[F_x(I - P) - F_x(I - P)A^{-1}A(I - P)] = 0.$$

Similarly, by (4.4b), (4.9), and (4.10)

$$\begin{aligned}\Psi^* F_x P &= \psi^* W_2 Q F_x [P - (I-P)A^{-1}(I-Q)F_x P] \\ &= \psi^* W_2 Q F_x \Omega_\varepsilon V_2^{-1} \\ &= \psi^* h_\varepsilon V_2^{-1} \\ &= 0.\end{aligned}$$

Therefore (4.7b) is satisfied.

The converse result, namely that (4.7) implies (4.6), can be obtained from (4.9) and (4.10) in a straightforward manner. We omit the details. Finally, we are left with showing (4.5), that is any null vector of $F_x(\Omega(\varepsilon, \alpha), \alpha)$ can be written in one of the forms in (4.4). Suppose $\Phi_1 \in N[F_x(\Omega(\varepsilon, \alpha), \alpha)]$ for (ε, α) near $(0, 0)$. Then set $\phi_1 = V^{-1}P\Phi_1$. Moreover, we have $0 = (I-Q)F_x(\Omega(\varepsilon, \alpha), \alpha)[(I-P)\Phi_1 + P\Phi_1]$ and therefore $(I-P)\Phi_1 = -A^{-1}(\varepsilon, \alpha)(I-Q)F_x V_2 \phi_1$.

Now using (4.10), we obtain

$$\Phi_1 = P\Phi_1 + (I-P)\Phi_1 = \Omega_\varepsilon(\varepsilon, \alpha)\phi_1.$$

Similarly, if $\Psi_1^* \in N[(F_x(\Omega(\varepsilon, \alpha), \alpha))^*]$ then it can be shown that $\Psi_1^* = \Psi^*(\varepsilon, \alpha, \psi_1)$ where $\psi_1 = W_2 Q \Psi_1$. □

From Lemma 4.4 we see that the assumption that $l = 1$ implies $h_\varepsilon(0, 0)$ has a one-dimensional null space, with right and left null vectors ϕ_0, ψ_0 such that $\Phi_0 = \Phi(0, 0, \phi_0)$, $\Psi_0 = \Psi(0, 0, \psi_0)$. However, the zero eigenvalue of h_ε^0 need not be simple, that is, it is possible that $\psi_0^* \phi_0 = 0$. But the eigenstructure of h_ε^0 depends on the choice of basis elements used to define V_2 and W_2 in (2.11). Indeed we note from (4.9c) and (4.10) that

$$h_\varepsilon(\varepsilon, \alpha) = W_2[QF_x\{I - A^{-1}(I-Q)\}]V_2.$$

Therefore, by reordering the basis vectors used to define V_2 and W_2, the rows and columns of h_ε can be rearranged. Similarly, by changing the sign of a basis vector, the sign of a row or column of h_ε can be reversed. The effect on ψ_0 and ϕ_0 is, of course, that their elements can be permuted and reversed in sign. By choosing an appropriate rearrangement the condition

$$\psi_0^* \phi_0 \neq 0 \tag{4.11}$$

can be obtained. (For example, change all negative coefficients in ψ_0 and ϕ_0 to positive coefficients, and then move the largest elements to the first position in each vector. Then (4.11) is a consequence of ψ_0 and ϕ_0 being nonzero.) Finally, we note that when (4.11) is satisfied, zero is a simple eigenvalue of h_ε^0 and therefore there exist smooth functions $\phi(\varepsilon, \alpha)$, $\psi(\varepsilon, \alpha)$, and $\bar{\mu}(\varepsilon, \alpha)$ such that

$$h_\varepsilon \phi = \bar{\mu}\phi, \ \psi^* h_\varepsilon = \mu \psi^* \tag{4.12a}$$

for (ε, α) near $(0, 0)$ with

$$\phi(0, 0) = \phi_0, \ \psi(0, 0) = \psi_0, \tag{4.12b}$$

$$\bar{\mu}(0, 0) = 0. \tag{4.12c}$$

We can now state the main result of this section.

THEOREM 4.13. *Suppose* $l = 1$ *and* $\Psi_0^* \Phi_0 \neq 0$. *Also, assume that* V_2 *and* W_2 *are such that* (4.11) *is satisfied. Then, for* (ε, α) *near* $(0, 0)$,

$$\mu(\varepsilon, \alpha) = a(\varepsilon, \alpha)\bar{\mu}(\varepsilon, \alpha) \tag{4.13a}$$

for some smooth function $a(\varepsilon, \alpha)$. *Moreover,*

$$\text{sign}\{a(0,0)\} = \text{sign}\{(\psi_0^* \phi_0)(\Psi_0^* \Phi_0)\}. \tag{4.13b}$$

Before proving Theorem 4.13 we remark that in the case $m = 1$ the result can be rewritten as

$$\mu(\varepsilon, \alpha) = a(\varepsilon, \alpha) h_\varepsilon(\varepsilon, \alpha)$$

where

$$\text{sign}\{a(0,0)\} = \text{sign}\{\Psi_0^* \Phi_0\}.$$

This is the result obtained in [7] for the Lyapunov-Schmidt decomposition. In particular, it provides the desired stability information for (ε, α) near $(0,0)$ in terms of a reduced function obtained from the generalised reduction procedure. For the case $m > 1$ the theorem illustrates how the stability information can be obtained from the behavior of the eigenvalue of $h_\varepsilon(\varepsilon, \alpha)$ that passes through zero at $(\varepsilon, \alpha) = 0$. In both cases the application of the theorem is trivial.

Proof of Theorem 4.13. Our proof is a simple modification of the proof of Theorem 4.1 presented in [7, p. 38]. It is clear from Lemma 4.4 that

$$\bar{\mu}(\varepsilon, \alpha) = 0 \quad \text{implies } \mu(\varepsilon, \alpha) = 0. \tag{4.14}$$

If $\nabla \mu^0 \equiv \partial \mu(0,0)/\partial(\varepsilon, \alpha) \neq 0$ and $\nabla \bar{\mu}^0 \neq 0$ then Proposition I.4.2 of [7] guarantees (4.13a) with a smooth (nonzero) $a(\varepsilon, \alpha)$.

Unfortunately, $\nabla \mu^0$ or $\nabla \bar{\mu}^0$ could be zero, in which case the proposition does not apply directly. It is convenient to add a new unfolding parameter to $F(x, \alpha)$ to ensure the applicability of Proposition I.4.2. of [7].

In particular, consider

$$\tilde{F}(x, \alpha, \beta) \equiv F(x, \alpha) + \beta \Psi_0 \Phi_0^* x = 0. \tag{4.15}$$

Then the same generalised reduction can be applied to (4.15). In particular, define $\tilde{\Omega}(\varepsilon, \alpha, \beta)$ to be the solution of (cf. (4.8a), (4.8b).

$$(I - Q)\tilde{F}(\tilde{\Omega}(\varepsilon, \alpha, \beta), \alpha, \beta) = 0, \tag{4.16a}$$

$$P\tilde{\Omega}(\varepsilon, \alpha, \beta) = V_2 \varepsilon, \tag{4.16b}$$

for $(\varepsilon, \alpha, \beta)$ near $(0,0,0)$. Then the reduced function is given by

$$\tilde{h}(\varepsilon, \alpha, \beta) \equiv W_2 Q \tilde{F}(\tilde{\Omega}(\varepsilon, \alpha, \beta), \alpha, \beta). \tag{4.17}$$

Note that this construction implies

$$h(\varepsilon, \alpha) = \tilde{h}(\varepsilon, \alpha, 0). \tag{4.18}$$

With some abuse of notation we extend μ, $\bar{\mu}$, Φ, Ψ, to be functions of $(\varepsilon, \alpha, \beta)$.

Next consider $\partial \mu / \partial \beta \ (0,0,0) \equiv \mu_\beta^0$. By differentiating

$$\tilde{F}_x(\tilde{\Omega}(\varepsilon, \alpha, \beta), \alpha, \beta)\Phi = \mu \Phi$$

with respect to β and evaluating at $(0,0,0)$, we find

$$\tilde{F}_{xx}^0 \Phi_0 \tilde{\Omega}_\beta^0 + \tilde{F}_{x\beta}^0 \Phi_0 + \tilde{F}_x^0 \Phi^0 = \mu_\beta^0 \Phi_0. \tag{4.19}$$

However, it follows from (4.16) that $\tilde{\Omega}_\beta^0 \equiv 0$. Furthermore, from (4.15) we have $\tilde{F}_{x\beta}^0 = \Psi_0 \Phi_0^*$. When we use these facts in (4.19), and apply Ψ_0^*, we are provided with

$$(\Psi_0^* \Psi_0)(\Phi_0^* \Phi_0) = \mu_\beta^0 \Psi_0^* \Phi_0. \tag{4.20}$$

In particular, $\mu_\beta^0 \neq 0$ and $\text{sign}\,(\mu_\beta^0) = \text{sign}\,(\psi_0^* \Phi_0)$.

A similar calculation based on

$$\tilde{h}_\varepsilon \phi = \bar{\mu}\phi$$

shows that

$$\bar{\mu}^0_\beta \psi_0^\top h^0_{\varepsilon\beta}\phi_0. \tag{4.21}$$

However, it follows from (4.17) and $\tilde{\Omega}^0_\beta = 0$ that

$$\tilde{h}^0_{\varepsilon\beta} = W_2 Q\{\tilde{F}^0_{x\beta}\tilde{\Omega}^0_\varepsilon + \tilde{F}^0_x \tilde{\Omega}^0_{\varepsilon\beta}\}.$$

By differentiating (4.16) twice we find

$$A^0(I-P)\tilde{\Omega}^0_{\varepsilon\beta} = -(I-Q)\tilde{F}^0_{x\beta}\tilde{\Omega}^0_\varepsilon, \qquad P\tilde{\Omega}^0_{\varepsilon\beta=0}.$$

By solving this expression for $\tilde{\Omega}^0_{\varepsilon\beta}$ and substituting the result into the above formula for $h^0_{\varepsilon\beta}$ we obtain

$$\tilde{h}^0_{\varepsilon\beta} = W_2 Q[I - F^0_x(I-P)A^{-1}(I-Q)]\tilde{F}^0_{x\beta}\Omega^0_\varepsilon.$$

Here we have used $\tilde{F}^0_x = F^0_x$, $A = A^0$, and $\tilde{\Omega}^0_\varepsilon = \Omega^0_\varepsilon$.

Finally, from $\tilde{F}^0_{x\beta} = \Psi_0 \Phi_0^*$ and (4.4), we have

$$\psi_0 \tilde{h}^0_{\varepsilon\beta}\phi_0 = \Psi_0^* \tilde{F}^0_{x\beta}\Phi_0 = (\Psi_0^*\Psi_0)(\Phi_0^*\Phi_0),$$

and so (4.21) becomes

$$\bar{\mu}^0_\beta = (\Psi_0^*\Psi_0)(\Phi_0^*\Phi_0)/(\psi_0^*\phi_0) \neq 0. \tag{4.22}$$

Now Proposition 4.2 of [7, § 4, Chap. I] ensures the existence of a smooth function $a(\varepsilon, \alpha, \beta)$ for $(\varepsilon, \alpha, \beta)$ near $(0, 0, 0)$ such that

$$\mu(\varepsilon, \alpha, \beta) = a(\varepsilon, \alpha, \beta)\bar{\mu}(\varepsilon, \alpha, \beta). \tag{4.23}$$

By differentiating (4.23) and using (4.20), (4.22) we find

$$a(0,0,0) = \bar{\mu}^0_\beta / \mu^0_\beta = \frac{\psi_0^*\phi_0}{(\Psi_0^*\Phi_0)} \neq 0. \tag{4.24}$$

The theorem follows now from (4.24) by setting $\beta = 0$ in (4.23). □

Appendix by A. Vanderbauwhede. In [7] Golubitsky and Schaeffer give a proof of the (contact) equivalence of the bifurcation functions that we obtain for different choices of the projections in a Lyapunov-Schmidt reduction. Here we give an alternative proof that, in our view, illustrates quite well the essential ideal of the Lyapunov-Schmidt reduction.

In the notation of § 2, suppose that the projections P and Q satisfy (2.5) with

$$X_2 = N[F^0_x], \qquad Y_1 = R[F^0_x].$$

It follows that (2.6) is satisfied, and therefore there is a local solution, $x_1 = \hat{v}(x_2, \alpha)$, to (2.7a). In fact, we find it more convenient to define $v: X_2 \times Y_1 \times R^k \to X_2$, where $k \equiv p+1$, and $v(u, w, \alpha)$ is the local solution of

$$(I-Q)F(u+v, \alpha) = w. \tag{A.1}$$

Clearly $v(u, 0, \alpha) = \hat{v}(u, \alpha)$. Substitution of this solution into (2.7b) provides the *bifurcation mapping* $g: X_2 \times R^k \to R[Q]$ defined by

$$g(u, \alpha) \equiv QF(u + v(u, 0, \alpha), \alpha). \tag{A.2}$$

For P and Q as above, this reduction provides one version of the Lyapunov-Schmidt procedure. The bifurcation mapping depends on the choice of projections P and Q; however, as the following theorem shows, different choices lead to equivalent mappings.

THEOREM. *Let $g: X_2 \times R^k \to R[Q]$ and $g_1: X_2 \times R^k \to R[Q_1]$ be two bifurcation mappings, obtained by choosing projections (P, Q) and (P_1, Q_1), respectively, in the reduction described above. Then there exist smooth mappings*

$$T: X_2 \times R^k \to L(R[Q_1], R[Q]),$$
$$U: X_2 \times R^k \to X_2,$$

such that

$$g(u, \alpha) = T(u, \alpha) g_1(U(u, \alpha), \alpha), \tag{A3a}$$

$$T(0, 0) = Q|_{R[Q_1]}, \tag{A3b}$$

$$U(0, 0) = 0 \quad \textit{and} \quad U_u(0, 0) = I_{X_2}. \tag{A3c}$$

Remark. All mappings in the statement and in the proof that follows are defined and smooth in a neighbourhood of the origin. The result shows that g and g_1 are contact equivalent at the origin (see [7]).

Proof. With $v(u, w, \alpha)$ defined as above, we set $V: X_2 \times Y_1 \times R^k \to X \times R^k$ to be

$$V(u, w, \alpha) \equiv (u + v(u, w, \alpha), \alpha). \tag{A4}$$

Then we have

$$V(0, 0, 0) = (0, 0), \tag{A5a}$$

$$\frac{\partial V}{\partial u}(0, 0, 0) = I_{X_2}, \quad \text{and} \quad F_x^0 \frac{\partial V}{\partial w}(0, 0, 0) = I_{Y_1}. \tag{A5b}$$

Therefore, we easily see that V is a local diffeomorphism. Moreover, from (A1), we have

$$F(V(u, w, \alpha)) = w + G(u, w, \alpha), \tag{A6}$$

for $G: X_2 \times Y_1 \times R^k \to R[Q]$ defined by

$$G(u, w, \alpha) \equiv QF(V(u, w, \alpha)). \tag{A7}$$

It is clear from the construction that

$$g(u, \alpha) = G(u, 0, \alpha). \tag{A8}$$

Now, by replacing (P, Q) by (P_1, Q_1) we find a similar local diffeomorphism $V_1: X_2 \times Y_1 \times R^k \to X \times R^k$ and a mapping $G_1: X_2 \times Y_1 \times R^k \to R[Q_1]$ such that

$$F(V_1(u, w, \alpha)) = w + G_1(u, w, \alpha), \tag{A9}$$

with

$$G_1(u, w, \alpha) \equiv Q_1 F(V_1(u, w, \alpha)). \tag{A10}$$

A second bifurcation mapping is given by

$$g_1(u, \alpha) = G_1(u, 0, \alpha). \tag{A11}$$

Our task is to show that g and g_1 are contact equivalent.

It is convenient to introduce $D: X_2 \times Y_1 \times R^k \to X_2 \times Y_1 \times R^k$ defined by $D \equiv V_1^{-1} \circ V$. It follows that D is a local diffeomorphism with $D(0, 0, 0) = (0, 0, 0)$. Furthermore, we define U and W by

$$D(u, 0, \alpha) = (U(u, \alpha), W(u, \alpha), \alpha). \tag{A12}$$

Upon differentiating $V_1 \circ D = V$ with respect to u and evaluating the result at $(0, 0, 0)$, we find from (A5b) that

$$U_u(0, 0) = I_{X_2}. \tag{A13}$$

We are now left with showing that (A3a) and (A3b) are satisfied for this definition of $U(u, \alpha)$.

From (A6) and (A9) it follows that

$$\begin{aligned} g(u, \alpha) &= F(V_1 \circ D(u, 0, \alpha)) \\ &= W(u, \alpha) + G_1(U(u, \alpha), W(u, \alpha), \alpha). \end{aligned} \tag{A14}$$

But (A11) and a first-order Taylor expansion together imply

$$G_1(u, w, \alpha) = g_1(u, \alpha) + T_1(u, w, \alpha) w \tag{A15}$$

for some smooth $T_1 : X_2 \times Y_1 \times R^k \to L(Y_1, R[Q_1])$. Since $\partial G_1 / \partial w\,(0, 0, 0) = 0$ we also have

$$T_1(0, 0, 0) = 0. \tag{A16}$$

Combining (A12), (A14), and (A15), we obtain

$$g(u, \alpha) = W(u, \alpha) + g_1(U(u, \alpha), \alpha) + T_1(D(u, 0, \alpha))\, W(u, \alpha).$$

Applying Q and $(I - Q)$ to this equation gives

$$\begin{aligned} g(u, \alpha) &= Q g_1(U, \alpha) + Q T_1(D)\, W, \\ (I_{Y_1} + (I - Q) T_1(D))\, W &= -(I - Q) g_1(U, \alpha), \end{aligned}$$

for $U = U(u, \alpha)$, $W = W(u, \alpha)$, and $D = D(u, 0, \alpha)$. Solving the second equation for W, and substituting the result into the first equation, provides (A3a) with

$$T(u, \alpha) = Q|_{R[Q_1]} - Q T_1(D)(I_{Y_1} + (I - Q) T_1(D))^{-1} (I - Q)|_{R[Q_1]}.$$

Equation (A3b) now follows from (A16). □

The central idea of the Lyapunov-Schmidt reduction is clearly revealed in (A6). Since V is a local diffeomorphism, we see from (A4) and (A5) that $F(x, \alpha) = 0$ is contact equivalent to

$$w + G(u, w, \alpha) = 0. \tag{A17}$$

Applying Q and $(I - Q)$ to this equation, we find that (A17) is equivalent to solving the reduced equation $g(u, \alpha) \equiv G(u, 0, \alpha) = 0$.

Acknowledgment. The authors thank Martin Golubitsky for providing an early draft of [7].

REFERENCES

[1] S. BANCROFT, J. K. HALE, AND D. SWEET, *Alternative problems for nonlinear functional equations*, J. Differential Equations, 4 (1968), pp. 40-56.

[2] L. BAUER, H. B. KELLER, AND E. L. REISS, *Multiple eigenvalues lead to secondary bifurcation*, SIAM Rev., 17 (1975), pp. 101-122.

[3] T. B. BENJAMIN AND T. MULLIN, *Anomalous modes in the Taylor Experiment*, Proc. Roy. Soc. London Ser. A, 359 (1981), pp. 24-43.

[4] W.-J. BEYN, *Defining equations for singular solutions and numerical applications*, Internat. Ser. Numer. Math., 70, pp. 42-56.

[5] L. CESARI, *Functional analysis and periodic solutions of nonlinear differential equations*, in Contributions to Differential Equations 1, 1963, pp. 149–187.

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

[7] ———*Singularities and Groups in Bifurcation Theory: Vol.* I, Applied Mathematical Sciences 51, Springer-Verlag, New York, 1985.

[8] A. D. JEPSON AND A. SPENCE, *Singular points and their composition*, Internat. Ser. Numer. Math., pp. 195–209.

[9] ———*The numerical solution of nonlinear equations having several parameters. Part* I: *Scalar equations*, SIAM J. Numer. Anal., 22 (1985), pp. 736–759.

[10] ———*The numerical solution of nonlinear equations having several parameters. Part* II: *Vector equations* submitted.

[11] M. A. KRASNOSEL'SKII, G. M. VAINIKKOO, P. P. ZABREIKO, YA. B. RUTITSKII, AND V. YA. STETSENKO, *Approximate Solution of Operator Equations*, Wolfers-Noordhoff, Groningen, 1972.

[12] M. SCHECHTER, *Principles of Functional Analysis*, Academic Press, New York, 1971.

[13] I. STACKGOLD, *Branching of solutions of nonlinear equations*, SIAM Rev., 13 (1971), pp. 289–332.

[14] I. STEWART, *Applications of nonelementary catastrophe theory*, IEEE Trans. Circuits and Systems, 31 (1984), pp. 165–174.

[15] M. M. VAINBERG AND T. A. TRENOGIN, *The methods of Liapunov and Schmidt in the theory of nonlinear equations and their further development*, Russian Math. Surveys, 17 (1962), pp. 1–60.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

005

VORTEX RINGS WITH SWIRL: AXISYMMETRIC SOLUTIONS OF THE EULER EQUATIONS WITH NONZERO HELICITY*

BRUCE TURKINGTON†

Abstract. This work introduces a new class of steady solutions of the axisymmetric Euler equations for an incompressible inviscid fluid. Each solution represents a three-dimensional vortex flow whose azimuthal components of vorticity and velocity are nonzero inside a toroidal region determined by the solution. The governing free-boundary problem is solved by variational techniques. The underlying variational principle is formulated from the natural invariants associated with the evolution equations for axisymmetric flows, and involves a family of invariants that generalizes the standard angular impulse and helicity integrals. A direct method is employed to prove the existence of steady solutions in a bounded domain and steadily translating solutions in space. Qualitative properties of these vortices are discussed and concentrated vortex rings with large swirl are shown to constitute a desingularization of the classical circular vortex filament.

Key words. Euler fluid dynamical equations, vortex, helicity, variational methods, free-boundary problems

AMS(MOS) subject classifications. 76C05, 49H05

Introduction. In this paper we examine a new class of steady solutions of the Euler equations governing the motion of an ideal fluid in three dimensions. The solutions that we consider are axisymmetric: they define flows that are invariant under rotation about the z-axis, when expressed in cylindrical coordinates (z, r, θ). Furthermore, each solution has the property that there is a (solid) toroidal region inside of which the θ-components of velocity u^θ and of vorticity ω^θ are nonzero, and outside of which the flow is irrotational. Therefore, we shall refer to these solutions, and the flows that they represent, as "vortex rings with swirl."

The objectives of our work are, first, to establish the existence of these steady solutions in a general and physically natural setting, and second, to derive some of their qualitative and asymptotic properties. We obtain our results by appealing to a variational formulation of the free-boundary problem satisfied by these solutions. The variational principle underlying our analysis of steady flows follows directly from the structure of the evolution equations governing the dynamics of axisymmetric vortex flows, which we express as a nonlinear, nonlocal system of equations for the quantities $\zeta = \omega^\theta / r$ and $\gamma = ru^\theta$. It is based on the reformulation of the steady equations in terms of ζ alone, which is accomplished easily through the elimination of γ. In this way we obtain a constrained maximization problem in ζ, the solutions of which define the desired steady vortex rings with swirl, and which is readily treated analytically and numerically. We utilize this approach to study both steady solutions in a bounded (axisymmetric) domain and steadily translating solutions in all of space.

An alternate variational characterization to ours has been given by Arnold [1], who considers fully three-dimensional flows, as well as by Benjamin [4], who specializes it to axisymmetric flows expressed in terms of ζ and γ. In this approach the variational principle is derived from the (noncanonical) Hamiltonian structure of the equations governing ζ and γ, and is formulated in the class of so-called isovortical variations of a given (extremal) flow. The resulting variational problem, though it arises very naturally

* Received by the editors October 1, 1986; accepted for publication (in revised form) May 2, 1988. The work of this author was partially supported by National Science Foundation DMS-8501795.

† Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003.

from the dynamical problem, is, unfortunately, highly nonconvex, and so is not amenable to the standard methods of analysis. Our variational problem, on the other hand, is designed to avoid such difficulties; instead, we take a wider class of competing functions, and compensate for this by adding to the objective (energy) functional certain terms that are invariant under all isovortical variations. These additional terms are constructed from the angular impulse (momentum) and helicity integrals. In fact, the construction involves a family of integrals generalizing the classical helicity integral; consequently, we are lead to introduce a family of conserved quantities—certain functionals of ζ and γ that we call "generalized angular impulse" and "generalized helicity integrals"—valid for evolving axisymmetric flows, which have not been used before in the literature. In turn, the variational principle that we give serves to clarify the role played by these generalized conserved quantities in determining the nature of steady axisymmetric flows.

The outline of the paper is as follows. In § 1, we discuss the axisymmetric Euler equations and the various conserved quantities associated with these equations. We begin § 2 by reviewing the general variational principles that characterize steady flows, and we conclude it by formulating the specific constrained maximization problem that we use in the further analysis. Section 3 is devoted to proving our main existence theorems for vortex rings with swirl. Finally, in § 4 we summarize some qualitative and asymptotic properties of the solutions found in § 3. In particular, we indicate how the parameters defining the angular impulse and helicity integrals determine the structure of the vortex ring and the flow field within it, and we identify the salient features of the solutions corresponding to extreme values of those parameters.

In a sequel to this paper [8] Eydeland and Turkington study propagating vortex rings with swirl in free space using an iterative numerical method allied with the variational structure of the governing problem formulated herein. The reader is referred there for a further exposition of these particularly interesting vortices and for a full discussion of their quantitative properties and physical characteristics.

1. Evolution equations and their invariants. Let (z, r, θ) denote the usual cylindrical coordinates in $\mathbf{R}^3$. Let $\hat{D} \subseteq \mathbf{R}^3$ be an axisymmetric domain, the axis of symmetry being $r = 0$; in the sequel, $\hat{D}$ will be either a bounded domain with a smooth boundary or all of space. We consider the axisymmetric flow of an ideal fluid with unit density in the domain $\hat{D}$, and we write the velocity and pressure fields in the form

$$u = u^z(z, r, t)e_z + u^r(z, r, t)e_r + u^\theta(z, r, t)e_\theta, \qquad p = p(z, r, t)$$

where $\{e_z, e_r, e_\theta\}$ is the usual coordinate frame. The governing equations are standard (see [2], for instance):

$$\begin{aligned} &u_t^z + u \cdot \nabla u^z = -p_z, \\ &u_t^r + u \cdot \nabla u^r - \frac{1}{r}(u^\theta)^2 = -p_r, \\ &u_t^\theta + u \cdot \nabla u^\theta + \frac{1}{r} u^r u^\theta = 0, \end{aligned} \tag{1.1}$$

$$u_z^z + \frac{1}{r}(ru^r)_r = 0, \tag{1.2}$$

where $u \cdot \nabla = u^z(\partial/\partial z) + u^r(\partial/\partial r)$ in view of the axisymmetry. For simplicity we will

impose the standard boundary conditions

$$\hat{v} \cdot u = 0 \quad \text{on } \partial\hat{D}, \tag{1.3}$$

where $\hat{v}$ is the outer unit normal on $\partial\hat{D}$.

The vorticity field, $\omega = \nabla \times u$, which plays a basic role in the subsequent development, is given by

$$\begin{aligned} \omega &= \omega^z(z, r, t)e_z + \omega^r(z, r, t)e_r + \omega^\theta(z, r, t)e_\theta \\ &= \frac{1}{r}(ru^\theta)_r e_z - u_z^\theta e_r + (u_z^r - u_r^z)e_\theta. \end{aligned} \tag{1.4}$$

The dynamical equations (1.1) are most conveniently expressed as evolution equations for the modified azimuthal vorticity, $\zeta = \zeta(z, r, t)$, and the azimuthal circulation (density), $\gamma = \gamma(z, r, t)$, which are defined by

$$\zeta = \frac{1}{r}\omega^\theta, \qquad \gamma = ru^\theta. \tag{1.5}$$

The equation for ζ is derived by taking the curl of the first two component equations in (1.1):

$$\begin{aligned} 0 &= \left[u_t^r + u \cdot \nabla u^r - \frac{1}{r}(u^\theta)^2\right]_z - [u_t^z + u \cdot \nabla u^z]_r \\ &= r\left[\left(\frac{\omega^\theta}{r}\right)_t + u \cdot \nabla\left(\frac{\omega^\theta}{r}\right) - 2r^{-2}u^\theta u_z^\theta\right]; \end{aligned} \tag{1.6}$$

thus, we obtain the first evolution equation

$$\zeta_t + u \cdot \nabla\zeta - 2r^{-4}\gamma\gamma_z = 0. \tag{1.7}$$

The equation for γ is equivalent to the conservation of circulation (Kelvin theorem), and it results from a manipulation of the third component equation in (1.1):

$$0 = u_t^\theta + u \cdot \nabla u^\theta + r^{-1}u^r u^\theta = r^{-1}[(ru^\theta)_t + u \cdot \nabla(ru^\theta)];$$

thus, we obtain the second evolution equation

$$\gamma_t + u \cdot \nabla\gamma = 0. \tag{1.8}$$

The continuity equation (1.2) furnishes a Stokes streamfunction $\psi = \psi(z, r, t)$ satisfying $u^z = \psi_r/r$ and $u^r = -\psi_z/r$. Consequently, ψ is determined by ζ alone according to

$$\zeta = r^{-1}[u_z^r - u_r^z] = L\psi$$

where L is the linear elliptic operator

$$L = -\frac{1}{r^2}\frac{\partial^2}{\partial z^2} - \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}\right). \tag{1.9}$$

In terms of ψ, the boundary condition (1.3) becomes $\psi = \text{const.}$ on $\partial\hat{D}$; we will assume that $\partial\hat{D}$ is connected, and that $\psi = 0$ on $\partial\hat{D}$.

We now introduce a new notation which simplifies the various formulas encountered in the rest of the paper. Let D denote the cross section of the spatial domain $\hat{D}$ in a meridional plane so that $\hat{D} = \{(z, r, \theta) \in \mathbf{R}^3 : (z, r) \in D\}$, and let $x = z$, $y = r^2/2$ be new (spatial) variables in D. Then, $\partial/\partial x = \partial/\partial z$, $\partial/\partial y = (1/r)\partial/\partial r$, and the volume

element dv (in $\hat{D}$) is replaced by $2\pi\, dx\, dy$. Also, the Jacobian of any two functions ϕ, ψ transforms according to $\partial(\phi, \psi)/\partial(x, y) = r^{-1}\partial(\phi, \psi)/\partial(z, r)$; we will abbreviate this expression to $\partial(\phi, \psi)$ throughout the sequel. The elliptic operator L is now given by

$$L = -\frac{1}{2y}\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}. \tag{1.10}$$

We let G denote the Green operator for L in D (with Dirichlet boundary conditions), so that $\psi = G\zeta$ defines the solution of

$$L\psi = \zeta \quad \text{in } D, \qquad \psi = 0 \quad \text{on } \partial D. \tag{1.11}$$

In this notation we may now write the governing equations (1.1)–(1.3) as a system of nonlinear, nonlocal evolution equations for ζ and γ:

$$\zeta_t + \partial(\zeta, G\zeta) + \partial(\gamma, \gamma/2y) = 0 \tag{1.12}$$

$$\gamma_t + \partial(\gamma, G\zeta) = 0 \qquad \text{in } D \times (0, T). \tag{1.13}$$

This form of the axisymmetric Euler equations is also given in [4].

The construction of the operator G merits further comment in view of the apparent singularity of L at $y = 0$. In the axisymmetric domain $\hat{D} \subseteq \mathbf{R}^3$ (which may contain a portion of the axis $r = 0$), let axisymmetric vector fields $\hat{\zeta}$ and $\hat{\psi}$ corresponding to the functions ζ and ψ be defined by $\hat{\zeta} := r\zeta e_\theta$ and $\hat{\psi} := (\psi/r)e_\theta$. Then, using $\nabla \cdot \hat{\psi} = 0$, we find that $L\psi = \zeta$ in D is equivalent to $-\Delta\hat{\psi} = \nabla \times \nabla \times \hat{\psi} = \hat{\zeta}$ in $\hat{D}$. Thus, the Green operator G takes L^2 into H_0^1 in the sense that the relevant norms are:

$$\int\int_D 2y\zeta^2\, dx\, dy = \int_{\hat{D}} |\hat{\zeta}|^2 \frac{dv}{2\pi} < +\infty, \qquad \int\int_D \left[\frac{1}{2y}\psi_x^2 + \psi_y^2\right] dx\, dy = \int_{\hat{D}} |\nabla\hat{\psi}|^2 \frac{dv}{2\pi} < +\infty.$$

Furthermore, since $\hat{\psi} \in H_0^1(\hat{D}) \cap H^2(\hat{D}) \subseteq C(\hat{D} \cup \partial\hat{D})$ it follows that $\psi = 0$ on ∂D and, by axisymmetry, that $\psi = 0$ at $y = 0$. (More details about this construction appear in the proofs of Theorems 1 and 2.)

In general, the existence and uniqueness of solutions of the initial value problem for (1.12), (1.13) can be asserted only on a sufficiently small time interval $0 \leqq t < T$; this result is just a special case of the known theory for fully three-dimensional flows. Thus, in our discussion of these equations we will proceed (formally) by assuming that a classical solution exists on some time interval $0 \leqq t < T$. Of course, when γ is identically zero (axisymmetric flow *without* swirl), equation (1.12) is solvable globally in time [15].

We now turn our attention to the conserved quantities associated with the evolution equations for ζ and γ. In this discussion we impose an additional boundary condition, namely,

$$\gamma = 0 \quad \textit{on } \partial D; \tag{1.14}$$

we note that if (1.14) holds at $t = 0$, then it holds also for all $t > 0$, by (1.13). With this (nonessential) condition in force, the derivation of the conserved quantities is simplified. Furthermore, all of the steady solutions that we construct in § 3 satisfy (1.14), and the purpose of the discussion here is to motivate the formulation of the main results of that section.

We claim that the following functionals are constant in the evolution governed by (1.12) and (1.13):

$$A(\gamma)=\iint_D a(\gamma)\,dx\,dy \qquad \text{(generalized angular impulse)}, \tag{1.15}$$

$$B(\zeta,\gamma)=\iint_D \zeta b(\gamma)\,dx\,dy \qquad \text{(generalized helicity)}, \tag{1.16}$$

$$C(\gamma)=\iint_D \zeta\,dx\,dy \qquad \text{(meridional circulation)}, \tag{1.17}$$

$$H(\zeta,\gamma)=\frac{1}{2}\iint_D\left[\zeta G\zeta+\frac{\gamma^2}{2y}\right]dx\,dy \qquad \text{(kinetic energy)}, \tag{1.18}$$

where a and b are arbitrary (suitably smooth) real functions. We leave to the reader the verification of the claims, which is easily accomplished with the aid of the integration by parts formula

$$\iint_D \phi_1\partial(\phi_2,\psi)\,dx\,dy=-\iint_D \phi_2\partial(\phi_1,\psi)\,dx\,dy,$$

which holds whenever either $\psi=\text{const.}$ or $\phi_1=0$ or $\phi_2=0$ on ∂D.

The classical cases of (1.15) and (1.16) occur when $a(\gamma)=\gamma$ and $b(\gamma)=\gamma$, since then we have

$$2\pi\iint_D \gamma\,dx\,dy=-\frac{1}{2}\int_{\hat{D}} r^2\omega^z\,dv, \qquad 2\pi\iint_D \zeta\gamma\,dx\,dy=\frac{1}{2}\int_{\hat{D}}\omega\cdot u\,dv.$$

We identify these integrals with the z-component of angular impulse (see [2]) and the helicity (see [12]), respectively, which are known invariants even for fully three-dimensional flows. The *generalized* angular impulse integral $A(\gamma)$ and helicity integral $B(\zeta,\gamma)$, valid for arbitrary functions a and b, seem to be new, however. The circulation integral

$$\iint_D \zeta\,dx\,dy=\iint_D \omega^\theta\,dz\,dr$$

has a standard interpretation by the Stokes theorem (see [2]). Upon integration by parts, the kinetic energy functional is recognized as

$$\pi\iint_D\left[\zeta G\zeta+\frac{\gamma^2}{2y}\right]dx\,dy=\frac{1}{2}\int_{\hat{D}}|u|^2\,dv.$$

The system of equations (1.12), (1.13) has a (noncanonical) Hamiltonian (or Poisson) structure, as has been noted in [4]. The functional $H(\zeta,\gamma)$ defined in (1.18) is the Hamiltonian, and a Poisson bracket $\{\cdot,\cdot\}$ can be defined so that $dF/dt=\{F,H\}$ for all (suitably smooth) functionals $F=F(\zeta,\gamma)$ defined on the appropriate phase space. Naturally, the functionals $F=A, B, C$ defined by (1.15)–(1.17) satisfy $\{F,H\}=0$, as can easily be verified.

2. Variational problems. The discussion in this section is divided into two parts. First, we present the fundamental variational principle that characterizes steady solutions of (1.12) and (1.13) directly in terms of the dynamical quantities ζ and γ. Second, we derive another version of this variational principle that involves only ζ (γ being eliminated algebraically), and that, when appropriately normalized, forms the basis of the subsequent analysis. This line of development is intended to motivate the specific variational problem that we employ in §§ 3 and 4 to study vortex rings with swirl.

We characterize a steady solution pair $\zeta = \zeta(x, y)$, $\gamma = \gamma(x, y)$ of (1.12), (1.13) variationally as follows. The governing equations become

$$\partial(\zeta, G\zeta) + \partial(\gamma, \gamma/2\gamma) = 0 \tag{2.1}$$

$$\partial(\gamma, G\zeta) = 0 \qquad \text{in } D. \tag{2.2}$$

Thus, to solve (2.2) we set

$$G\zeta = b(\gamma), \tag{2.3}$$

where b is a specified (suitably smooth) function with $b(0) = 0$ (recalling (1.14)). Substitution of (2.3) into (2.1) then yields

$$0 = \partial(\zeta, b(\gamma)) + \partial(\gamma, \gamma/2y) = \partial(\zeta b'(\gamma) - \gamma/2y, \gamma).$$

Thus, to solve this equation we set

$$\zeta b'(\gamma) - \gamma/2y = -a'(\gamma), \tag{2.4}$$

where a is a specified (suitably smooth) function. Therefore, given (essentially arbitrary) a and b, it suffices to solve (2.3) and (2.4) for the desired pair ζ, γ. We now define the modified energy functional (assuming that appropriate dimensional constants scale the given functions a and b):

$$\tilde{H}(\zeta, \gamma) = H(\zeta, \gamma) - A(\gamma) - B(\zeta, \gamma) \tag{2.5}$$

where the specified functions a, b determine the functionals A, B in (1.15), (1.16). Then it is obvious that (2.3) and (2.4) are equivalent to $\tilde{H}_\zeta = 0$ and $\tilde{H}_\gamma = 0$, respectively, where $\tilde{H}_\zeta$ and $\tilde{H}_\gamma$ denote the Fréchet derivatives of $\tilde{H}$ with respect to ζ and γ. In this straightforward manner, we arrive at the variational principle:

(VP1) Any critical point (ζ, γ) for the functional $\tilde{H}(\zeta, \gamma)$ is a solution of (2.1), (2.2), and hence yields a steady flow in $\hat{D}$.

An alternate variational principle given in [4] (and implicit in [1]) can be summarized as follows. For arbitrary test functions ϕ_1, ϕ_2 (each ϕ_i is smooth in $\bar{D}$ and vanishes on ∂D), consider the variations $\tilde{\zeta} = \tilde{\zeta}(x, y; s)$, $\tilde{\gamma} = \tilde{\gamma}(x, y; s)$ defined by solving the equations

$$\tilde{\zeta}_s + \partial(\tilde{\zeta}, \phi_1) + \partial(\tilde{\gamma}, \phi_2) = 0, \qquad \tilde{\gamma}_s + \partial(\tilde{\gamma}, \phi_1) = 0,$$

with $\tilde{\zeta}|_{s=0} = \zeta$, $\tilde{\gamma}|_{s=0} = \gamma$. Then (2.1), (2.2) result from the variational equation

$$\frac{d}{ds} H(\tilde{\zeta}, \tilde{\gamma})|_{s=0} = 0 \quad \text{for arbitrary } \phi_1, \phi_2.$$

As is remarked in [4], this characterization of solutions of (2.1), (2.2) is quite cumbersome to use in analysis because of the particular form of the class of variations involved. In contrast, (VP1) permits arbitrary variations by replacing H with $\tilde{H} = H - A - B$. In this regard, we note that $A(\gamma)$, $B(\zeta, \gamma)$, and $C(\gamma)$ are invariant under the variations $\tilde{\zeta}$, $\tilde{\gamma}$ defined above, as is easily checked.

We now reduce (VP1) to a more special variational principle that does not involve γ explicitly. For this reduction we require that

$$a(0)=0, \quad b(0)=0, \quad a'(t)\leqq 0, \quad b'(t)>0 \quad \text{for all } t\geqq 0, \tag{2.6}$$

and we define

$$f(y, s) := [b^{-1}(s)]^2/4y - a(b^{-1}(s)), \qquad s\geqq 0. \tag{2.7}$$

Then we see that a pair ζ, γ satisfies (2.1), (2.2) whenever γ is defined by $\gamma = b^{-1}(G\zeta)$ and ζ satisfies the equation

$$\zeta = f_s(y, G\zeta). \tag{2.8}$$

We now assume further that a and b are such that $f(y, s)$ is strictly convex in s (this holds in the "classical" case when both a and b are linear, for instance). Let $f^*(y, \sigma)$, the conjugate function to $f(y, s)$, be defined by

$$f^*(y, \sigma) = \sup_s [s\sigma - f(y, s)]. \tag{2.9}$$

Then f^* has the well-known properties

$$f^*(y, \sigma) = sf_s(y, s) - f(y, s), \qquad f^*_{\sigma\sigma}(y, \sigma) = [f_{ss}(y, s)]^{-1}$$

with $\sigma = f_s(y, s)$ or, equivalently, $s = f^*_\sigma(y, \sigma)$. Equation (2.7) may be rewritten in this notation as

$$G\zeta = f^*_\sigma(y, \zeta), \tag{2.10}$$

and this is clearly the variational equation for the functional

$$\Phi(\zeta) = \int\int_D \left[\frac{1}{2}\zeta G\zeta - f^*(y, \zeta)\right] dx\, dy. \tag{2.11}$$

Consequently, we have the variational principle:

(VP2) Any extremal ζ for the functional $\Phi(\zeta)$ yields a solution of (2.1), (2.2) with $\gamma = b^{-1}(G\zeta)$.

The algebraic elimination of γ in terms of ζ may also be viewed as a restriction on the admissible variations in (VP1). As is readily verified using the properties of the convex conjugate function, the expression for γ implied by (2.4) is

$$\gamma = \Gamma(\zeta) := b^{-1}(f^*_\sigma(y, \zeta)); \tag{2.12}$$

indeed, this inverts the equation (equivalent to (2.4))

$$\zeta = \gamma/2yb'(\gamma) - a'(\gamma)/b'(\gamma).$$

We now claim that the identity (2.12) reduces the objective functional in (VP1) as follows:

$$\Phi(\zeta) = \tilde{H}(\zeta, \Gamma(\zeta)) \quad \text{for arbitrary admissible } \zeta. \tag{2.13}$$

To check this it suffices to observe that

$$-f^*(y, \sigma) = \frac{[b^{-1}(s)]^2}{4y} - a(b^{-1}(s)) - \frac{s}{b'(b^{-1}(s))}\left\{\frac{b^{-1}(s)}{2y} - a'(b^{-1}(s))\right\},$$

and, consequently, that

$$\begin{aligned} f^*(y, \zeta) &= \frac{\gamma^2}{4y} - a(\gamma) - \frac{b(\gamma)}{b'(\gamma)}\left\{\frac{\gamma}{2y} - a'(\gamma)\right\} \\ &= \frac{\gamma^2}{4y} - a(\gamma) - \zeta b(\gamma). \end{aligned}$$

On the basis of the foregoing discussion, we now formulate the specific constrained maximization problems that we solve in § 3 to obtain steady and steadily translating vortex rings with swirl. These problems involve free boundaries, since the solutions have the property that $\zeta, \gamma > 0$ in a subdomain $\Omega \subset D$ and $\zeta, \gamma = 0$ in $D \backslash \Omega$. Therefore, we introduce some additional constraints and normalizations into (VP2). We assume that $a(t)$ and $b(t)$ are specified satisfying (2.6) and that $f_{ss}(y, s) > 0$. We consider the problem:

(2.14) maximize $\Phi(\zeta)$ subject to the constraints $\zeta \geqq 0$ in D, $C(\zeta) := \iint_D \zeta \, dx \, dy = C_0$,

where C_0 is a (suitably specified) positive constant. We claim that such a maximizer ζ yields a solution of (2.1), (2.2) having the form

$$\begin{aligned} &\zeta = f_s(y, G\zeta - \mu), \quad \gamma = b^{-1}(G\zeta - \mu) \quad \text{in } \Omega := \{G\zeta > \mu\}, \\ &\zeta = 0, \quad \gamma = 0 \quad \text{in } D \backslash \Omega, \end{aligned} \tag{2.15}$$

where μ is the Lagrange multiplier for the constraint $C(\zeta) = C_0$. To verify the claim we calculate the variational conditions at a maximizer ζ; we get

$$\Phi'(\zeta) - \mu = 0 \quad \text{on } \{\zeta > 0\}, \qquad \Phi'(\zeta) - \mu \leqq 0 \quad \text{on } \{\zeta = 0\},$$

where $\Phi'(\zeta) = G\zeta - f_\sigma^*(y, \zeta)$ is the Fréchet derivative of Φ at ζ. These conditions imply the claimed expression for ζ, since $\{0 < \zeta < f_s(y, 0)\} \subseteq \{G\zeta = \mu\}$, and hence, invoking the argument of Corollary 2.3 in [13], meas $\{0 < \zeta < f_s(y, 0)\} = 0$. (If $f_s(y, 0) = 0$, then ζ is continuous across $\partial\Omega$.) The claimed expression for γ is immediate from (2.12). That ζ and γ given by (2.15) satisfy (2.1) and (2.2) follows exactly as in the deviation of (VP2). Thus, we have the desired variational characterization of steady vortex rings with swirl in a bounded domain $\hat{D}$.

We formulate the constrained maximization problem for steadily translating vortex rings with swirl in $\mathbf{R}^3$ in the same way. In this case the objective functional $\Phi(\zeta)$ (defined by (2.11) with $\mathbf{R}^2_+$ replacing D) is maximized subject to the constraints

$$\zeta \geqq 0 \quad \text{in} \quad \mathbf{R}^2_+, \quad C(\zeta) := \iint_{\mathbf{R}^2_+} \zeta \, dx \, dy = C_0, \quad P(\zeta) := \iint_{\mathbf{R}^2_+} y\zeta \, dx \, dy = P_0. \tag{2.16}$$

The maximizer ζ then yields a solution of (2.1), (2.2) having the form

$$\begin{aligned} &\zeta = f_s(y, G\zeta - cy - \mu), \quad \gamma = b^{-1}(G\zeta - cy - \mu) \quad \text{in } \Omega := \{G\zeta > cy + \mu\}, \\ &\zeta = 0, \quad \gamma = 0 \quad \text{in } \mathbf{R}^2_+ \backslash \Omega, \end{aligned} \tag{2.17}$$

where c and μ are the Lagrange multipliers for the constraints $P(\zeta) = P_0$ and $C(\zeta) = C_0$, respectively. The additional constraint, $P(\zeta) = P_0$ fixes the z-component of linear impulse (see [2]), a further conserved quantity in this case given by

$$2\pi \iint_{\mathbf{R}^2_+} y\zeta \, dx \, dy = \frac{1}{2} \int_{R^3} r\omega^\theta \, dv.$$

The constant c represents the translational velocity, and is determined by the solution.

It is important to note that both the problem in D with constraints (2.14) and the problem in $\mathbf{R}^2_+$ with constraints (2.16) can be naturally nondimensionalized by appropriately normalizing their constraint values. Indeed, if L and U are characteristic length and velocity scales, respectively, then the constraint values scale according to $C_0 = ULC_0^*$ and $P_0 = UL^4P_0^*$ when expressed in terms of dimensionless variables (indicated by stars): $x = Lx^*$, $y = L^2y^*$, $\zeta = UL^{-2}\zeta^*$, etc. Therefore, throughout the sequel we

will assume that $C_0 = 1$ and diameter $(D) = 1$ (say) when dealing with solutions in a bounded domain $\hat{D}$, and that $C_0 = 1$ and $P_0 = 1$ when dealing with solutions in all of space $\mathbf{R}^3$.

We remark that the steady flows considered above can also be characterized in terms of their streamfunctions ψ, and a variational problem for ψ, which is dual to (VP2), can be given. However, we prefer the equivalent formulation in ζ because (i) it is more closely tied to governing dynamical equations (through (VP1)), (ii) the constraints imposed in (2.14) and (2.16) are more natural physically than a specification of μ and c, and (iii) the asymptotic properties of concentrated vortex rings with swirl are more readily obtained (see § 4). These advantages have been utilized in the theory of steady vortex rings without swirl in [3] and [11].

3. Existence theorems. In this section we state and prove two existence theorems for vortex rings with swirl: Theorem 1 concerns steady solutions in a bounded (axisymmetric) domain $\hat{D} \subset \mathbf{R}^3$, while Theorem 2 concerns steadily translating solutions in $\mathbf{R}^3$. We establish both of these theorems by applying direct variational methods to the corresponding constrained maximization problems formulated in § 2. This approach is quite standard and has been used before in similar problems. The results obtained in [5] parallel those in our Theorem 1, but unlike that paper we treat the case when the domain $\hat{D}$ contains a portion of the z-axis (where certain singularities arise). Also, the results of our Theorem 2 are analogous to those established in [11], although the approach we take in the present paper is more direct.

For the sake of simplicity in the exposition we will restrict our detailed discussion in this section to the "classical" case when

$$a(t) = -\alpha t, \quad b(t) = \beta t \quad \text{for given positive constants } \alpha \text{ and } \beta. \tag{3.1}$$

This special case illustrates well the general case: some instances of the general case for which the same results hold are discussed later in this section.

In the first theorem the constrained maximization problem under consideration is defined on the class of competing functions

$$K(D) = \left\{ \zeta \in L^1(D) : \zeta \geqq 0 \text{ a.e.}, \iint_D \zeta \, dx\, dy \leqq 1, \iint_D y\zeta^2 \, dx\, dy < +\infty \right\}. \tag{3.2}$$

The constraint $C(\zeta) = 1$ is relaxed to an inequality here for technical reasons related to the range of the given parameters α and β. We recall that the objective functional introduced in § 2, specialized according to (3.1), is defined as

$$\Phi(\zeta) = \iint_D \left[\frac{1}{2} \zeta G\zeta - y\beta^2 \left(\zeta - \frac{\alpha}{\beta} \right)_+^2 \right] dx\, dy. \tag{3.3}$$

We let χ_Ω denote the characteristic function of $\Omega \subseteq D$.

THEOREM 1. *For every prescribed $0 \leqq \alpha < +\infty$, $0 < \beta < +\infty$ there exists a solution $\zeta \in K(D)$ of the problem*

$$\text{maximize } \Phi(\zeta) \text{ subject to } \zeta \in K(D), \tag{3.4}$$

and there exists a multiplier $\mu > 0$ such that

$$\zeta = \frac{1}{2y\beta^2} (G\zeta - \mu)_+ + \frac{\alpha}{\beta} \chi_{\{G\zeta - \mu > 0\}} \text{ in } D. \tag{3.5}$$

Furthermore, whenever $\mu > 0$ there holds

$$\zeta \text{ has compact support in } D, \ C(\zeta) = 1. \tag{3.6}$$

Remark. The condition that $\mu > 0$ is ensured whenever β is small enough depending on α, and this property is proved in § 4. The exact range of the parameter β, an interval $0 < \beta < \beta^*(\alpha)$, for which solutions satisfying (3.6) exist is best determined numerically using the method given in [8].

Proof. We first construct a maximizer for Φ over $K(D)$. To be able to treat the case when $\hat{D} \cap \{r = 0\} \neq \varnothing$, it is necessary to reexpress the problem in terms of the (z, r, θ) coordinates and the naturally associated vector fields:

$$\hat{\zeta} = r\zeta(z, r)e_\theta, \quad \hat{\psi} = \frac{1}{r}\psi(z, r)e_\theta \quad \text{in } D. \tag{3.7}$$

We verify that $L\psi = \zeta$ in D with $\psi = 0$ on ∂D if and only if $-\Delta\hat{\psi} = \hat{\zeta}$ in $\hat{D}$ with $\hat{\psi} = 0$ on $\partial\hat{D}$; we see this equivalence by virtue of the identity $-\Delta\hat{\psi} = \nabla \times \nabla \times \hat{\psi}$, since $\nabla \cdot \hat{\psi} = 0$. Now the terms in the objective functional Φ can be expressed in $\hat{\zeta}$ and $\hat{\psi}$. In particular, we have

$$\frac{1}{2}\int\int_D \zeta G\zeta \, dx\, dy = \frac{1}{2}\int_{\hat{D}} \hat{\zeta} \cdot \hat{\psi} \, dm = \frac{1}{2}\int_{\hat{D}} |\nabla\hat{\psi}|^2 \, dm, \qquad \int\int_D y\zeta^2 \, dx\, dy = \frac{1}{2}\int_{\hat{D}} |\hat{\zeta}|^2 \, dm,$$

where we write $dm = r\, dr\, dz = (1/2\pi)\, dv$. Also, we have

$$\int\int_D |\hat{\zeta}| \, dm \leqq R := \max_{\hat{D}} r,$$

by the circulation constraint. An upper bound for Φ on $K(D)$ is obtained as follows. By the Sobolev inequality, we have

$$\int_{\hat{D}} |\nabla\hat{\psi}|^2 \, dm = \int_{\hat{D}} \hat{\zeta} \cdot \hat{\psi} \, dm \leqq \|\hat{\zeta}\|_{6/5}\|\hat{\psi}\|_6 \leqq C_1\|\hat{\zeta}\|_{6/5}\left\{\int_{\hat{D}} |\nabla\hat{\psi}|^2 \, dm\right\}^{1/2}.$$

This yields the estimate

$$\left\{\int_{\hat{D}} |\nabla\hat{\psi}|^2 \, dm\right\}^{1/2} \leqq C_2\|\hat{\zeta}\|_{6/5} \leqq C_2\|\hat{\zeta}\|_1^{2/3}\|\hat{\zeta}\|_2^{1/3} \leqq C_3(R)\|\hat{\zeta}\|_2^{1/3},$$

on application of the standard interpolation inequality. We now obtain the desired bound for Φ: for $\zeta \in K(D)$,

$$\begin{aligned} \Phi(\zeta) &\leqq \frac{1}{2}\int_{\hat{D}} |\nabla\hat{\psi}|^2 \, dm - \frac{\beta^2}{2}\int_{\hat{D}} |\hat{\zeta}|^2 \, dm + \alpha\beta \int_{\hat{D}} r|\hat{\zeta}| \, dm \\ &\leqq C_4(R)\left\{\int_{\hat{D}} |\hat{\zeta}|^2 \, dm\right\}^{1/3} - \frac{\beta^2}{2}\int_{\hat{D}} |\hat{\zeta}|^2 \, dm + \alpha\beta R^2 \\ &\leqq C_5(D, \alpha, \beta) < +\infty. \end{aligned} \tag{3.8}$$

Consequently, we may take a sequence $\zeta_j \in K(D)$ such that (i) $\Phi(\zeta_j) \to \sup\{\Phi(\tilde{\zeta}): \tilde{\zeta} \in K(D)\}$, (ii) $\|\hat{\zeta}_j\|_2 \leqq C_6$, and (iii) $\hat{\zeta}_j \to \hat{\zeta} \in L^2(\hat{D})$ weakly. It follows that $\hat{\psi}_j \to \hat{\psi}$ weakly in $H^2(\hat{D})$ and hence, by the standard imbedding theorem, that $\hat{\psi}_j \to \hat{\psi}$ strongly in $H_0^1(\hat{D})$. Thus, we have the continuity of the first term of Φ, namely,

$$\int\int_D \zeta_j G\zeta_j \, dx\, dy = \int_{\hat{D}} |\nabla\hat{\psi}_j|^2 \, dm \to \int_{\hat{D}} |\nabla\hat{\psi}|^2 \, dm = \int\int_D \zeta G\zeta \, dx\, dy.$$

Also, we have the lower semicontinuity of the second term, namely,

$$\liminf_{j\to+\infty} \int\int_D y\left(\zeta_j - \frac{\alpha}{\beta}\right)_+^2 dx\, dy \geqq \int\int_D y\left(\zeta - \frac{\alpha}{\beta}\right)_+^2 dx\, dy.$$

This follows from the convexity of this term combined with the identity $2y\zeta = re_\theta \cdot \hat{\zeta}$, which implies that $\zeta_j \to \zeta$ weakly in $L^2(D)$ with respect to the measure $y\,dx\,dy$. Consequently, we may conclude that $\Phi(\zeta) = \lim_{j\to+\infty} \Phi(\zeta_j) = \sup \Phi$, with $\zeta \in K(D)$. (We note that the statement that limits $\hat{\zeta}$ and $\hat{\psi}$ have the form (3.7) is easily checked.)

The derivation of the variational conditions (3.5) is sketched in § 2 in the general case. In the present case, if μ is the multiplier accounting for the constraint $C(\zeta) \leqq 1$, then we have, by maximality,

$$0 \geqq \int\int_D [G\zeta - \mu - 2y\beta^2(\zeta - \alpha/\beta)_+]\delta\zeta\,dx\,dy$$

for all variations $\delta\zeta$ that respect the constraint $\zeta \geqq 0$ almost everywhere in D. This inequality implies that

$$\begin{aligned} &G\zeta - \mu = 2y\beta^2(\zeta - \alpha/\beta)_+ \quad \text{whenever } \zeta > 0, \\ &G\zeta - \mu \leqq 0 \qquad\qquad\qquad\quad \text{whenever } \zeta = 0, \end{aligned} \tag{3.9}$$

which further implies that $\{0 < \zeta < \alpha/\beta\} \subseteq \{G\zeta = \mu\}$. The stated form (3.5) then follows immediately.

To complete the proof we must establish (3.6) when $\mu > 0$, by assumption. Since $\hat{\zeta} \in L^2(\hat{D})$, we have that $\hat{\psi} \in H_0^1(\hat{D}) \cap H^2(\hat{D}) \subseteq C^{1/2}(\hat{D} \cup \partial\hat{D})$, by the Morrey-Sobolev imbedding theorem. Then we must have $\hat{\psi}|_{r=0} = 0$, by continuity, and hence $\psi = r|\hat{\psi}| = o(r)$ as $r \to 0^+$. Now since $\mu > 0$, it follows that supp $\zeta \subseteq \{\psi \geqq \mu\} \subseteq D \cap \{y > \delta\}$ for some $\delta > 0$, and thus we conclude that supp ζ is a compact subset of D. The fact that the corresponding multiplier is nonzero ensures that equality holds in the circulation constraint.

The above theorem provides a solution pair ζ, γ of the system (2.1), (2.2) in the "classical" case (3.1), and that solution pair is expressed as

$$\zeta = \frac{1}{2y\beta^2}\tilde{\psi}_+ + \frac{\alpha}{\beta}\chi_{\{\tilde{\psi}>0\}}, \quad \gamma = \frac{1}{\beta}\tilde{\psi}_+ \quad \text{with } \tilde{\psi} = G\zeta - \mu. \tag{3.10}$$

The theorem generalizes in a straightforward way to a class of problems where the structure functions $a, b \in C^2[0, +\infty)$ satisfy (2.6) and $f(y, s)$ defined in (2.7) is strictly convex in s for $s > 0$. In particular, it suffices that $a(t)$ and $b(t)$ satisfy

$$a(0) = 0, \quad -\alpha_1 \leqq a'(t) \leqq 0, \quad a''(t)b'(t) \geqq a'(t)b''(t), \tag{3.11}$$

$$b(0) = 0, \quad \beta_0 \leqq b'(t) \leqq \beta_1, \quad b'(t) > tb''(t), \tag{3.12}$$

for all $t \geqq 0$, where $\alpha_1, \beta_0, \beta_1$ are positive constants. The solution pair ζ, γ is then expressible as

$$\begin{aligned} &\zeta = \frac{b^{-1}(\tilde{\psi})}{2yb'(b^{-1}(\tilde{\psi}))} - \frac{a'(b^{-1}(\tilde{\psi}))}{b'(b^{-1}(\tilde{\psi}))} \quad \text{in } \{\tilde{\psi} > 0\}, \quad \zeta = 0 \quad \text{in } \{\tilde{\psi} \leqq 0\}, \\ &\gamma = b^{-1}(\tilde{\psi}) \quad \text{in } \{\tilde{\psi} > 0\}, \quad \gamma = 0 \quad \text{in} \quad \{\tilde{\psi} \leqq 0\}. \end{aligned} \tag{3.13}$$

It follows from this that $\zeta \in L^\infty(D)$ and that ζ jumps by the constant $|a'(0)|/b'(0)$ across the free boundary $\partial\{\tilde{\psi} > 0\}$; it also follows that $\gamma \in C^{0,1}(D)$ and that, in general, γ_x and γ_y are discontinuous across the free boundary. In physical terms, the velocity field is continuous everywhere (and that condition defines the free-boundary condition), while the vorticity field may be discontinuous across the boundary of the vortex ring. Of course, further generalizations are certainly possible (for instance, the quadratic growth of $f(y, s)$ in s can be relaxed and solutions with higher regularity can be obtained), but these extensions are left to the reader.

In the second theorem, we replace the domain D by $\mathbf{R}^2_+ = \{y > 0\}$, and we impose the additional constraint $P(\zeta) = 1$. The class of competing functions is now taken to be

$$K(\mathbf{R}^2_+) = \Big\{ \zeta \in L^1(\mathbf{R}^2_+) : \zeta \geqq 0 \text{ a.e.}, \iint_{\mathbf{R}^2_+} \zeta \, dx\, dy \leqq 1, \tag{3.14}$$
$$\iint_{\mathbf{R}^2_+} y\zeta \, dx\, dy \leqq 1, \iint_{\mathbf{R}^2_+} y\zeta^2 \, dx\, dy < +\infty \Big\}.$$

Both of the constraints $C(\zeta) = 1$ and $P(\zeta) = 1$ are relaxed to inequalities here again for technical reasons. The proof of existence of solutions follows much as in the preceding theorem, except that it is complicated by the fact that the domain is now unbounded. □

THEOREM 2. *For every prescribed $0 \leqq \alpha < +\infty$, $0 < \beta < +\infty$, there exists a solution $\zeta \in K(\mathbf{R}^2_+)$ of the problem*

$$\text{maximize } \Phi(\zeta) \text{ subject to } \zeta \in K(\mathbf{R}^2_+) \tag{3.15}$$

and there exist multipliers $\mu \geqq 0$ and $c > 0$ such that

$$\zeta = \frac{1}{2y\beta^2}(G\zeta - cy - \mu)_+ + \frac{\alpha}{\beta}\chi_{\{G\zeta - cy - \mu > 0\}} \text{ in } \mathbf{R}^2_+; \tag{3.16}$$

also, ζ is symmetrized about $x = 0$, in the sense that

$$\zeta(x, y) = \zeta(-x, y) \text{ and } \zeta(x, y) \geqq \zeta(x', y) \text{ whenever } 0 \leqq x \leqq x'. \tag{3.17}$$

Furthermore, $P(\zeta) = 1$ holds and whenever $\mu > 0$ there also holds

$$\zeta \text{ has compact support in } \mathbf{R}^2_+, \ C(\zeta) = 1. \tag{3.18}$$

Remark. As in Theorem 1, the condition that $\mu > 0$ is ensured whenever β is small enough depending on α. In fact, the range $0 < \beta < \beta^*(\alpha)$ in Theorem 2 is known explicitly, since the limiting case $\beta = \beta^*(\alpha)$ corresponds to a spherical vortex with swirl found by Moffatt [12]. A complete discussion of these Moffatt vortices and the role they play in determining the range of the given parameters α and β is given in [8], where the vortex rings with swirl proven to exist in this theorem are exhibited numerically.

Proof. We use the same direct variational method as in the proof of Theorem 1, leaving some of the technical details to the reader. As before, we introduce $\hat{\zeta}$ and $\hat{\psi}$ defined by (3.7). However, the constraints $C(\zeta) \leqq 1$ and $P(\zeta) \leqq 1$ now imply the bound

$$\int_{\mathbf{R}^3} |\hat{\zeta}| \, dm \leqq \int_{\mathbf{R}^3} \frac{1}{2}(r + r^{-1})|\hat{\zeta}| \, dm = \iint_{\mathbf{R}^2_+} \left(y + \frac{1}{2}\right)\zeta \, dx\, dy \leqq \frac{3}{2}.$$

In turn, this yields the bound

$$\Phi(\zeta) \leqq C_4 \left\{ \int_{\mathbf{R}^3} |\hat{\zeta}|^2 \, dm \right\}^{1/3} - \frac{\beta}{2\lambda} \int_{\mathbf{R}^3} |\hat{\zeta}|^2 \, dm + 2\alpha\beta \leqq C_5(\alpha, \beta) < +\infty, \tag{3.19}$$

where, in contrast to (3.8), the constant C_4 is independent of the size of the support of $\hat{\zeta}$. It follows that there is a sequence $\zeta_j \in K(\mathbf{R}^2_+)$ such that (i) $\Phi(\zeta_j) \to \sup\{\Phi(\tilde{\zeta}) : \tilde{\zeta} \in K(\mathbf{R}^2_+)\}$, (ii) $\|\hat{\zeta}_j\|_2 \leqq C_6$, and (iii) $\hat{\zeta}_j \to \hat{\zeta} \in L^2(\mathbf{R}^3)$ weakly. Furthermore, using the standard arguments (see [9]), we may replace each $\zeta_j(x, y)$ by its symmetrical rearrangement in the x variable so that the properties (3.16) hold for each ζ_j; in this procedure we use the fact that symmetrization in x does not alter the constraints and

does not decrease the objective functional. Now, however, we are able to assert only that $\hat{\psi}_j \to \hat{\psi}$ strongly in $H^1(B_\rho)$, where $B_\rho = \{z^2 + r^2 < \rho^2\}$ for every fixed $\rho < +\infty$. Consequently, the proof of continuity of the first term of Φ_λ requires careful justification. This is provided by the estimate

$$|\hat{\psi}_j(z, r)| \leqq \eta(\sqrt{z^2 + r^2}) \quad \text{where } \eta(\rho) \downarrow 0 \quad \text{as } \rho \to +\infty. \tag{3.20}$$

Indeed, given such a rate of decay for $|\hat{\psi}_j|$ (independently of j) we conclude that

$$\iint_{z^2+r^2 \geqq \rho^2} \zeta_j G\zeta_j \, dx \, dy = \int_{\mathbf{R}^3 \setminus B_\rho} \hat{\zeta}_j \cdot \hat{\psi}_j \, dm \leqq \frac{3}{2} \eta(\rho),$$

and similarly for ζ; this combined with the fact that

$$\iint_{z^2+r^2<\rho^2} \zeta_j G\zeta_j \, dx \, dy = \int_{B_\rho} |\nabla \hat{\psi}_j|^2 \, dm \to \int_{B_\rho} |\nabla \hat{\psi}|^2 \, dm = \iint_{z^2+r^2<\rho^2} \zeta G\zeta \, dx \, dy$$

for arbitrarily large $\rho < +\infty$, clearly yields the desired conclusion. The proof of the claimed estimate (3.20) depends upon the fact that $\hat{\psi}_j$ is expressible as the potential of $\hat{\zeta}_j$, namely, $\hat{\psi}_j = k_* \hat{\zeta}_j$ (component-wise), where $k = k(\rho) = 1/4\pi\rho$ is the fundamental solution of $-\Delta$ in $\mathbf{R}^3$—and $|\hat{\zeta}_j|$ is axisymmetric, symmetrized in z, and $\|\hat{\zeta}_j\|_1 \leqq C_1$, $\|\hat{\zeta}_j\|_2 \leqq C_2$ for constants C_1, C_2 independent of j. A derivation of such an estimate can be given by modifying the calculations made in [10], where an analogous result is proved with the L^∞-bound $\|\hat{\zeta}_j\|_\infty \leqq C_2$ replacing the L^2-bound; the resulting estimate shown in Theorem 2.5 of [10] involves $\eta(\rho) = C_3 \rho^{-1} \log(1 + C_4 \rho)$, and an analogous expression follows in the present case. Since the details of this argument are rather lengthy and are of technical interest only, we will omit them here. The lower semicontinuity of the second term in Φ follows as before. Therefore, we have that $\Phi(\zeta) = \lim_{j \to \infty} \Phi(\zeta_j) = \sup_K \Phi$, with $\zeta \in K(\mathbf{R}^2_+)$.

The variational conditions (3.15) follow using the standard methods, with multipliers c, $\mu \in \mathbf{R}$ uniquely determined by the extremal ζ. That c, $\mu \geqq 0$ follows from the observation that $|\hat{\zeta}| = r\zeta = (\psi - cr/2 - \mu/r)^+ \geqq (-cr/2 - \mu/r)^+$ in $\mathbf{R}^3$ with $|\hat{\zeta}| \in L^1(\mathbf{R}^3)$. Thus, $c < 0$ implies $|\hat{\zeta}| \geqq |c| r/2 - |\mu|/r$ for large r, which is impossible; and, $\mu < 0$ implies $|\hat{\zeta}| \geqq |\mu|/r - |c| r/2$ for small r, which is impossible. The strict positivity of c is an immediate consequence of the following identity, which is interesting in itself:

$$2c = \iint_{\mathbf{R}^2_+} \left[\frac{1}{2} \zeta G\zeta + \beta^2 y \left(\zeta - \frac{\alpha}{\beta} \right)_+^2 + 6\alpha\beta y \left(\zeta - \frac{\alpha}{\beta} \right)_+ \right] dx \, dy. \tag{3.21}$$

To prove this we use an alternative expression for the energy associated with flow in the meridional planes, namely,

$$\frac{1}{2} \iint_{\mathbf{R}^2_+} \left[\frac{1}{2y} \psi_x^2 + \psi_y^2 \right] dx \, dy = \iint_{\mathbf{R}^2_+} (x\psi_x + 2y\psi_y) L\psi \, dx \, dy. \tag{3.22}$$

This formula is verified by an integration by parts argument; its complete derivation is given in Lemma 3.1 of [11]. Now we let $\tilde{\psi} = \psi - cy - \mu$, and using (3.22) we obtain the expression

$$\begin{aligned} \frac{1}{2} \iint_{\mathbf{R}^2_+} \zeta G\zeta \, dx \, dy &= \frac{1}{2} \iint_{\mathbf{R}^2_+} \left[\frac{1}{2y} \psi_x^2 + \psi_y^2 \right] dx \, dy \\ &= \iint_{\mathbf{R}^2_+} [x\tilde{\psi}_x + 2y(\tilde{\psi} + cy)_y] \zeta \, dx \, dy. \end{aligned}$$

But an integration by parts yields

$$\iint_{\mathbf{R}^2_+} x\tilde{\psi}_x\zeta\,dx\,dy = \iint_{\mathbf{R}^2_+} x\tilde{\psi}_x\left[\frac{1}{2y\beta^2}\tilde{\psi}_+ + \frac{\alpha}{\beta}\chi_{\{\tilde{\psi}>0\}}\right]dx\,dy$$

$$= -\iint_{\mathbf{R}^2_+}\left[\frac{1}{4y\beta^2}\tilde{\psi}_+^2 + \frac{\alpha}{\beta}\tilde{\psi}_+\right]dx\,dy,$$

$$\iint_{\mathbf{R}^2_+} 2y(\tilde{\psi}+cy)_y\zeta\,dx\,dy = \iint_{\mathbf{R}^2_+}\tilde{\psi}_y\left[\frac{1}{\beta^2}\tilde{\psi}_+ + \frac{2y\alpha}{\beta}\chi_{\{\tilde{\psi}>0\}}\right]dx\,dy$$

$$+2c\iint_{\mathbf{R}^2_+} y\zeta\,dx\,dy$$

$$= -\iint_{\mathbf{R}^2_+}\frac{2\alpha}{\beta}\tilde{\psi}_+\,dx\,dy + 2c.$$

The claimed identity (3.21) now follows using the substitution $\tilde{\psi}_+ = 2y\beta^2(\zeta - \alpha/\beta)_+$.

We note that equality must hold in the constraint $P(\zeta)\leqq 1$ since c is nonzero. It remains to establish (3.17) when $\mu > 0$, by assumption. First, we note, as before, that equality must hold in the constraint $C(\zeta)\leqq 1$. Next, we demonstrate that supp $\zeta \subseteq \{|x|\leqq\delta^{-1},\ \delta\leqq y\leqq\delta^{-1}\}$ for some $\delta > 0$. The required bounds in y follow directly from the fact that $\psi \geqq cy+\mu$ in supp ζ, since this implies that $cr/2+\mu r^{-1}\leqq|\hat{\psi}|\leqq C_6$ in supp ζ. The required bounds in x follow from the inequality

$$2|x|y \leqq c^{-2}\iint_{\mathbf{R}^2_+}\zeta G\zeta\,dx'\,dy' \quad \text{for all } (x,y)\in\text{supp }\zeta.$$

To show this we use the fact that $\psi(x',y)=\psi(-x',y)$ and $\psi(x',y)\geqq\psi(x,y)$ for all $0<x'<x$ (a consequence of (3.16)), because then we obtain

$$2c|x|y \leqq \int_{-|x|}^{|x|}\psi(x',y)\,dx' = \int_{-|x|}^{|x|}\int_0^y \psi_{y'}(x',y')\,dx'\,dy'$$

$$\leqq (2|x|y)^{1/2}\left\{\iint_{\mathbf{R}^2_+}\psi_{y'}^2\,dx'\,dy'\right\}^{1/2},$$

which clearly gives the desired inequality. This completes the proof. □

The above theorem supplies a solution pair ζ, γ of the system (2.1), (2.2) in the "classical" case (3.1), and that solution pair is expressed as

$$\zeta = \frac{1}{2y\beta^2}\tilde{\psi}_+ + \frac{\alpha}{\beta}\chi_{\{\tilde{\psi}>0\}}, \quad \gamma = \frac{1}{\beta}\tilde{\psi}_+ \quad \text{with } \tilde{\psi} = G\zeta - cy - \mu. \tag{3.23}$$

Remarks made after Theorem 2 concerning the generalization of the existence result to a class of structure functions a and b satisfying (3.11) apply equally well to the results of Theorem 2.

4. Qualitative properties of vortex rings with swirl. Here we summarize some of the qualitative properties of the solutions found in Theorems 1 and 2. The specific form of the constrained maximization problem we employ in the analysis of solutions provides an especially convenient framework for the derivation of these results, particularly those pertaining to asymptotic properties. However, we will be content merely to sketch the proofs of the results stated here, since our method of analysis has appeared in several papers [7], [11], [13], [14] concerning very similar problems.

As in § 3 and in our sequel paper [8], we restrict our attention to the "classical" case in which the structure functions a and b satisfy (3.1). Then it is of interest to determine how the solutions $\zeta = \zeta_{\alpha,\beta}$, $\gamma = \gamma_{\alpha,\beta}$ depend on the given parameters α and β. We will first consider the dependence upon β (for fixed α), emphasizing the asymptotic limit $\beta \to 0+$ that gives concentrated vortex rings with swirl. We will then consider the dependence of these β-parametrized branches of solutions on α, which dictates the flow structure within the vortex.

The asymptotic analysis of solutions as $\beta \to 0+$ can be accomplished with a modification of the methods given in [13]. The crucial estimates are as follows (C_1, C_2, etc. denote generic positive constants independent of β):

$$\text{diam}\,(\Omega) \leqq C_1\beta, \tag{4.1}$$

$$\max\,(G\zeta - \mu)_+ \leqq C_2, \quad \max\,\zeta \leqq C_3\beta^{-2}, \quad \max\,\gamma \leqq C_4\beta^{-1}. \tag{4.2}$$

The geometric estimate (4.1) on the vortex core $\Omega = \{\tilde{\psi} > 0\}$ may be derived using an extension of the arguments of Theorem 3.3 in [13]. We now briefly indicate the main steps in this derivation for the solutions found in Theorem 1. First, we note the identity

$$\mu = 2\Phi(\zeta) - 2\alpha\beta \iint_D y\left(\zeta - \frac{\alpha}{\beta}\right)_+ dx\,dy, \tag{4.3}$$

which follows immediately from (3.5). Second, we establish that

$$2\Phi(\zeta) \geqq 2\Phi(\hat{\zeta}) \geqq \frac{\hat{r}}{2\pi}\log\beta^{-1} - C_1,$$

taking an admissible function $\hat{\zeta} = (1/\pi\beta^2)\chi_{B_\beta(\hat{x},\hat{y})}$ approximating a delta function centered at some point $(\hat{x}, \hat{y}) \in D$. This estimate depends upon some precise information about the Green function for L in D; in particular, the Green function is known to have the same singular behavior as the fundamental solution for L in $\mathbf{R}^2_+$ (the stream-function for the classical circular vortex filament), which is given by

$$k(z, r, z', r') = \frac{(rr')^{1/2}}{2\pi}\log\xi^{-1} + O(1) \quad \text{as } \xi \to 0+$$

where $\xi = [(z-z')^2 + (r-r')^2]^{1/2}/2(rr')^{1/2}$, when expressed in terms of the variables $z = x$, $r = \sqrt{2y}$ (see [11]). It follows that for any $(x, y) \in \Omega$

$$G\zeta(x, y) \geqq \mu \geqq 2\Phi(\hat{\zeta}) - C_2 \geqq \frac{\hat{r}}{2\pi}\log\beta^{-1} - C_3.$$

A relatively crude argument using this estimate will demonstrate that diam $(\Omega) = o(1)$ as $\beta \to 0+$, and with this fact in hand (and taking $(\hat{x}, \hat{y}) \in \Omega$) the sharp version of the argument ([13, Thm. 3.3) yields the desired estimate (4.1). These arguments rely on showing that, as a consequence of the latter inequality, if $(x, y) \in \Omega$ then

$$\iint_{D \setminus B_{M\beta}(x,y)} \zeta(x', y')\,dx'\,dy' \leqq c_4(\log M)^{-1} \quad \text{for } M > 1,$$

and hence that diam $(\Omega) \leqq 2M\beta$, when M is fixed large enough that $C_4(\log M)^{-1} < \frac{1}{2}$. Once (4.1) is proved, the asymptotically sharp estimates (4.2) follow from the scaling arguments given in Theorems 4.4 and 4.5 of [13]. We omit any further details here.

An obvious consequence of the identity (4.3) is the positivity of μ for small β; in fact, $\mu \geqq a_0 \log\beta^{-1}$ for some constant $a_0 > 0$ as $\beta \to 0+$. Thus, we have justified the remark following Theorem 1, and hence equality must hold in the constraint $C(\zeta) \leqq 1$

for small β. (Presumably, μ is positive for some parameter range $0<\beta<\beta^*(\alpha)$, which could be determined numerically using the method in [8].)

We see from (4.1), (4.2) that ζ tends to a unit delta function in the sense of distributions as $\beta\to 0+$. (The location at which the delta function is concentrated can be deduced from the Green function for L in D, as is explained in Theorem 4.3 of [13].) Also, we see that γ tends to zero in the sense of distributions as $\beta\to 0+$, even though $\max \gamma\to+\infty$. This rather curious property is shared by all concentrated vortex rings with (nonzero) swirl. Consequently, the limit solution is a (typically unique) vortex filament (without swirl) in D which is independent of α. A further manifestation of this phenomenon is that the angular impulse $A(\gamma)\to 0$, while the helicity $B(\zeta,\gamma)=0(\beta^{-1})$ as $\beta\to 0+$.

Similar asymptotic results hold for the solutions found in Theorem 2, namely, steadily translating vortex rings with swirl in all of space. First, recalling (3.21), we observe that the translational speed c satisfies

$$2c=\Phi(\zeta)+\int\int_{\mathbf{R}^2_+}\left[2\beta^2 y\left(\zeta-\frac{\alpha}{\beta}\right)_+^2+6\alpha\beta y\left(\zeta-\frac{\alpha}{\beta}\right)_+\right]dx\,dy. \tag{4.4}$$

Then, since (4.1), (4.2) are also valid for these solutions, it follows that $c\geqq b_0\log\beta^{-1}$ for some constant $b_0>0$ as $\beta\to 0+$. The identity (4.3) now holds with μ replaced by $\mu+c$, and so, in turn, it follows that $\mu\geqq a_0\log\beta^{-1}$ for some constant $a_0>0$. (In fact, the following asymptotic formulas as $\beta\to 0+$ are easily established: $\Phi(\zeta)\sim(\sqrt{2}/4\pi)\log\beta^{-1}$, $c\sim\frac{1}{2}\Phi(\zeta)$, $\mu\sim\frac{3}{2}\Phi(\zeta)$.) As before, we see that as $\beta\to 0+$

$$\zeta\to\delta(x,y-1),\ \gamma\to 0 \text{ in the sense of distributions,}$$

where $\delta(x,y-1)$ is the unit delta function at $(x,y)=(0,1)$. In other words, the unique circular vortex filament with unit linear impulse is obtained as the limit solution, independently of α.

These results justify the remark following Theorem 2, since μ is positive for small β, and hence $C(\zeta)\leqq 1$ holds as an equality. However, much more precise information on the parameter range $0<\beta<\beta^*(a)$ over which this holds is available and is given in [8]. There it is shown that the extreme case $\beta=\beta^*(\alpha)$, for which $\mu=0$, corresponds to an explicit spherical vortex found by Moffatt [12], and the function $\beta^*(\alpha)$ is calculated. The reader is referred there for the details as well as for a quantitative description of the solutions over the full parameter range.

We now turn to the dependence of the solutions on the parameter α. While β determines the cross-sectional diameter of the vortex ring with swirl, α controls the structure of the vortical flow within the vortex ring. When $\alpha=0$, it is readily verified that (3.10) is equivalent to the identity $u=\beta\omega$ in the vortex core. Thus, every solution with $\alpha=0$ defines a vortex ring that consists of a Beltrami flow within its vortex core. Roughly speaking, these solutions represent steady vortex flows whose swirl is maximal with respect to their meridional circulation. (Unlike simple quasi-two-dimensional flows it is not possible for the swirl γ to take essentially arbitrary values, because of the presence of the coupling term $\partial(\gamma,\gamma/2y)$ in (1.12).) The opposite extreme occurs when $\alpha\to+\infty$. Since $\beta^*(\alpha)\to+\infty$ as $\alpha\to+\infty$ (as can be checked), there exist limit solutions $\hat{\zeta}_\lambda:=\lim\zeta_{\alpha,\beta}$ when $\alpha,\beta\to+\infty$ and $\alpha/\beta\to\lambda$ (provided that $\lambda>\lambda^*$, say). In the case of Theorem 1, these solutions satisfy

$$\hat{\zeta}_\lambda=\lambda\chi_{\{G\hat{\zeta}_\lambda>\mu\}},\quad \hat{\gamma}_\lambda=0 \quad \text{in } D,$$

which corresponds to the limit of the free-boundary problems (3.10) for finite α,β. Consequently, such a limit solution represents a classical steady vortex ring without

swirl (see [3], [10]). The relevant parameter range is $\lambda^* < \lambda < +\infty$ where $(\lambda^*)^{-1} = \iint_D dx\,dy$. Similar results apply in the case of Theorem 2, where a branch of steadily translating vortex rings is obtained for which $\lambda = \lambda^*$ gives the Hill spherical vortex and $\lambda = +\infty$ gives the circular vortex filament (both normalized by the constraints $C(\zeta) = 1$, $P(\zeta) = 1$). In either case the general situation with $0 < \alpha < +\infty$ may be viewed as mediating between the extremes described above. Hence the two-parameter family of solutions $\zeta_{\alpha,\beta}$, $\gamma_{\alpha,\beta}$ furnishes a very natural and precise extension of the familiar concept of a vortex ring.

We conclude by commenting that our theoretical results are tied to some interesting phenomena on the physical side. In an unpublished experiment, Bergerud [6] devised an apparatus for imparting azimuthal swirl to concentrated vortex rings, and observed that steady flows were produced when the swirl had a certain critical magnitude, while unsteady (azimuthally oscillating) flows were produced for other swirl magnitudes. It is noteworthy that the steady vortex rings with swirl he observed in water share some qualitative features with our model flows in an ideal fluid. In particular, the bound $A(\gamma) \leqq \beta P(\zeta)$, which holds for all of the solutions given in Theorem 2, confirms that the total angular impulse of any steadily translating vortex ring must be small if it is concentrated (β small) and has a given linear impulse ($P(\zeta)$ prescribed). However, further experimentation is needed before a convincing evaluation of the relevance of our solutions to vortex motions in real fluids can be made.

REFERENCES

[1] V. I. ARNOLD, *Variational principle for three-dimensional steady-state flows of an ideal fluid*, J. Appl. Math. Mech., 29 (1965), pp. 1002–1008.

[2] G. K. BATCHELOR, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, 1974.

[3] T. B. BENJAMIN, *The alliance of practical and analytical insights into the nonlinear problems of fluid mechanics*, in Applications of Methods of Functional Analysis to Problems in Mechanics, Lecture Notes in Mathematics 503, Springer-Verlag, Berlin, 1976, pp. 8–29.

[4] ———, *Impulse, flow force and variational principles*, IMA J. Appl. Math., 32 (1984), pp. 3–68.

[5] H. BERESTYCKI AND H. BREZIS, *On a free boundary problem arising in plasma physics*, Nonlinear Anal.: Theory, Methods Appl., 4 (1980), pp. 415–436.

[6] D. BERGERUD, personal communication.

[7] L. A. CAFFARELLI AND A. FRIEDMAN, *Asymptotic estimates for the plasma problem*, Duke Math. J., 47 (1980), pp. 705–742.

[8] A. EYDELAND AND B. TURKINGTON, *A numerical study of vortex rings with swirl*, J. Fluid Mech. (1988), to appear.

[9] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, Wiley-Interscience, New York, 1982.

[10] A. FRIEDMAN AND B. TURKINGTON, *Asymptotic estimates for an axisymmetric rotating fluid*, J. Funct. Anal., 37 (1980), pp. 136–163.

[11] ———, *Vortex rings: existence and asymptotic estimates*, Trans. Amer. Math. Soc., 268 (1981), pp. 1–37.

[12] H. K. MOFFATT, *The degree of knottedness of tangled vortex lines*, J. Fluid Mech., 35 (1969), pp. 117–129.

[13] B. TURKINGTON, *On steady vortex flow in two dimensions*, I, Commun. PDE, 8 (1983), pp. 999–1030.

[14] ———, *On steady vortex flow in two dimensions*, II, Commun. PDE, 8 (1983), pp. 1031–1071.

[15] V. I. YUDOVICH, *Non-stationary flow of an ideal incompressible fluid.* Zh. Vychisl. Mati. Mat. Fiz., 3 (1963), pp. 1032–1066. (In Russian.) USSR Comput. Math. and Math. Phys., 3 (1963), pp. 1407–1456. (In English.)

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

006

STATIONARY STOKES AND NAVIER–STOKES SYSTEMS ON TWO- OR THREE-DIMENSIONAL DOMAINS WITH CORNERS. PART I: LINEARIZED EQUATIONS*

MONIQUE DAUGE†

Abstract. The H^s-regularity (s being real and nonnegative) of solutions of the Stokes system in domains with corners is studied. In particular, a H^2-regularity result on a convex polyhedron that generalizes Kellogg and Osborn's result on a convex polygon to three-dimensional domains is stated. Sharper regularity on a cube and on other domains with corners is attained. Conditions for the problem to be Fredholm are also given, and its singular functions along with those of the nonlinear problem are studied in the second part of this paper.

Key words. Stokes, corner, edge, polyhedron, regularity of solutions

AMS(MOS) subject classifications. 35Q, 76N

1. Introduction. The linearized equations corresponding to the Navier–Stokes system describing gas-dynamics consist of the following Stokes system in $\mathbf{R}^n$ ($n=2$ or 3):

$$-\Delta \vec{u} + \nabla p = \vec{f}, \qquad \operatorname{div} \vec{u} = g \tag{1.1}$$

where $\vec{u} = (u_1, \cdots, u_n)$ is the speed of the fluid, p its pressure, and $\vec{f}$ the strength field. On a domain Ω, the boundary conditions are

$$\vec{u}_{|\partial\Omega} = 0. \tag{1.2}$$

The problem (1.1)–(1.2) can be approached as an elliptic boundary value problem as in the paper by Agmon, Douglis, and Nirenberg [1]. On the other hand, it may be proved by a variational method (see Temam [22]) that for a bounded domain Ω and data $(\vec{f}, g)$ in the product of Sobolev spaces $[H^{-1}(\Omega)]^n \times L^2(\Omega)$ with the compatibility condition

$$\int_\Omega g \, dX = 0, \tag{1.3}$$

there exists a unique solution $(\vec{u}, p)$ of (1.1)–(1.2) in the space $[\mathring{H}^1(\Omega)]^n \times [L^2(\Omega)/\mathbf{C}]$. Here, as usual, $\mathring{H}^1(\Omega)$ denotes the H^1-space with null traces on the boundary, and H^{-1} is its dual with respect to the L^2-duality.

Thus, if $(\vec{f}, g)$ is more regular, let us say

$$\vec{f} \in [H^{s-1}(\Omega)]^n \quad \text{and} \quad g \in H^s(\Omega), \quad s > 0, \tag{1.4}$$

then, when Ω has a smooth boundary, we draw from [1] and interpolation (cf. [23]), that

$$\vec{u} \in [H^{s+1}(\Omega)]^n \quad \text{and} \quad p \in H^s(\Omega). \tag{1.5}$$

But, in the case of physical domains, or for partition of domains in numerical analysis, it is natural to study the case when Ω has corners.

In two-dimensional domains (2D), when Ω is a polygon, we have Kondrat'ev's [12] and Grisvard's [10] results for the divergence-free system ($g=0$: incompressible fluid) in spaces with integer exponents; for the general system (1.1), we have Osborn's

* Received by the editors August 11, 1987; accepted for publication (in revised form) May 2, 1988.

† U.A. Centre National de la Recherche Scientifique 758, Département de Mathématiques et Informatique, 2, rue de la Houssinière, 44072 Nantes Cedex 03, France.

results [19], Dauge's results [5] in weighted Sobolev spaces, and the regularity result of Kellogg and Osborn [11].

In three-dimensional domains (3D), Maz'ja and Plamenevskii study the problem (1.1)-(1.2) for a large class of domains in weighted Sobolev spaces: the results are announced in [15] and proved in [16], [17a], [17b]. The spaces are general L^p-Sobolev spaces with weight (of Kondrat'ev type) and also Hölder classes with weight. Merigot [18] and Grisvard [10] have also used L^p-Sobolev spaces in the 2D divergence-free problem on a polygon.

In this paper we state precise results of regularity in the ordinary spaces (1.4), (1.5). Among other things, the Sobolev spaces with real exponents are useful for studying the nonlinear Navier-Stokes system (Part II of this work is forthcoming), and for successive approximation schemes (see [20]).

Theorems 5.4 and 5.5 in 2D are a generalization of [10] and [11]. In 3D we get new results. For several examples of domains, we hereafter indicate a condition on s under which the solution $(\vec{u}, p)$ of (1.1)-(1.2) with $(\vec{f}, g)$ in the space (1.4) has the regularity of (1.5), provided g is zero at the singular points of Ω if $s \geqq 1$ (cf. [11] and the definition (9.17)):

(1.6) If Ω is any domain in our class of domains with corners $\mathcal{O}_3$ (introduced in § 2 below), $s < 0.5$.

(1.7) If $\Omega = Q_1 \backslash Q_2$ where Q_1 and Q_2 are two rectangular parallelepipeds with the same axes, $s \leqq 0.544$ (approximate value).

(1.8) If Ω is any convex domain in our class $\mathcal{O}_3$, $s \leqq 1$.

(1.9) If Ω is any convex domain with wedge angles $\leqq 2\pi/3$, $s < 3/2$.

(1.10) If Ω is any cylinder with convex polygonal base, and angles $< 2\pi/3$, $s \leqq 3/2$.

(1.11) If Ω is any cylinder with smooth base, $s < 2$.

(1.12) If Ω is a half-ball, $s < 2$.

When we say a cylinder, we mean a bounded cylinder truncated perpendicularly to its generating lines.

The plan of this paper is as follows. In § 2 we introduce our classes of domains and the functional spaces. In § 3 we recall general results from Dauge's works [6] and [9], and we apply them to the problem (1.1)-(1.2). As these results are based on a special condition of injectivity about tangent problems, in § 4 we link that condition to the usual one used by Kondrat'ev in [12]. In § 5 we recall some properties of the characteristic equation $\sin^2 \lambda\omega - \lambda^2 \sin^2 \omega = 0$, we give a graph and tables of values for its roots, and we state results in 2D. In § 6 we study the domains in 3D that have edges, but no vertices. In § 7 we study the tangent problem in a three-dimensional cone, which gives rise to a quantity linked with the Laplace-Beltrami operator that we estimate in § 8. Finally, we state 3D results in § 9.

2. Classes of domains and functional spaces. Our classes of domains contain various curvilinear polygons (in 2D) and polyhedra or domains with piecewise-smooth boundary (in 3D).

2.1. Plane and spherical domains. Our class $\mathcal{O}_2(\mathbf{R}^2)$ of plane domains consists of all curvilinear polygons, possibly with cracks but without cusps (or turning points): Ω is in $\mathcal{O}_2(\mathbf{R}^2)$ if and only if it enjoys the following properties:

(i) Ω is bounded and connected.

(ii) The boundary of Ω consists of a finite number of smooth closed arcs $\Gamma_1, \cdots, \Gamma_N, \Gamma_{N+1} = \Gamma_1$.

(iii) Let A_j and A_{j+1} be the ends of Γ_j; the A_j for $j=1,\cdots,N$ are the vertices of Ω and at the neighborhood of A_j, Ω is locally diffeomorphic to a neighborhood of zero in a plane sector Γ_{A_j}.

In the case when one of the sectors Γ_{A_j} has its opening equal to 2π, we have a *crack* and we dissociate the two sides of the crack as in Fig. 1.

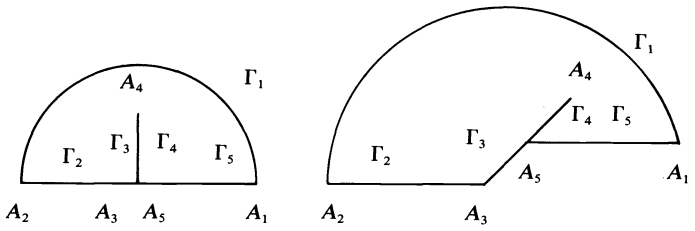


FIG. 1

In both cases, A_4 is at the bottom of the crack and Γ_3 and Γ_4 coincide in the neighborhood of A_4.

Let us note that condition (iii) may be rewritten in the following form. If the tangents of Γ_j and Γ_{j+1} coincide in A_j, then Γ_j and Γ_{j+1} coincide in a neighborhood of A_j.

We denote by $A_0(\Omega)$ the set $\{A_1,\cdots,A_N\}$ and denote simply by x any element of $A_0(\Omega)$. Thus, for $x=A_j$, Γ_{A_j} is denoted by Γ_x.

In the same way we define the class $\mathcal{O}_2(S^2)$ of curvilinear polygons on the unit sphere of $\mathbf{R}^3$.

2.2. Three-dimensional domains. Ω belongs to $\mathcal{O}_3(\mathbf{R}^3)$ if and only if it satisfies the following conditions:

(i) Ω is bounded and connected.

(ii) At each point x of its "stretched" boundary, Ω is locally diffeomorphic to a neighborhood of zero in one of the following three kinds of domains:

(1) A half-space: then x is a **regular** point;

(2) A dihedron isomorphic to $\mathbf{R}\times\Gamma_x$, with Γ_x a plane sector with an opening ω_x different from π: then x belongs to an **edge**;

(3) A cone Γ_x with vertex zero (which is not a dihedron), such that its intersection G_x with S^2 belongs to $\mathcal{O}_2(S^2)$: then x is a **vertex**.

Let $A_0(\Omega)$ be the set of vertices and $A_1(\Omega)$ be the union of the edges.

The **stretched boundary** is the notion corresponding to the doubling of the boundary when there is a crack in 2D. This is more completely explained in § 2 of [9].

Note that if Ω has a piecewise-smooth boundary, and its faces meet two by two or three by three with independent normals at meeting points, then Ω belongs to our class $\mathcal{O}_3(\mathbf{R}^3)$.

2.3. Sobolev spaces. For a positive integer s, $H^s(\Omega)$ is the usual Sobolev space of all distributions u in $\mathcal{D}'(\Omega)$ such that each derivative $D^\alpha u$ with length $|\alpha|\leqq s$, in Ω, belongs to $L^2(\Omega)$. For a positive noninteger real number s, let $[s]$ be the integer part of s and $\sigma=s-[s]$. $H^s(\Omega)$ is the space of all u in $H^{[s]}(\Omega)$ that satisfy

$$\forall\alpha,\ |\alpha|=[s]\quad \iint_\Omega |D^\alpha u(x)-D^\alpha u(y)|^2\, d(x,y)^{-n-2s}\, dx\, dy<+\infty$$

where $d(x,y)$ is the infimum of length of the paths joining x to y and included in Ω.

$\mathring{H}^s(\Omega)$ is the closure of $\mathscr{D}(\Omega)$ in $H^s(\Omega)$ and $H^{-s}(\Omega)$ its dual with respect to the L^2-duality.

2.4. Stokes operators. We denote by $D_n^s(\Omega)$ the product of Sobolev spaces $[\mathring{H}^1 \cap H^{s+1}(\Omega)]^n \times H^s(\Omega)$ (cf. (1.5)) and by $E_n^s(\Omega)$ the product $[H^{s-1}(\Omega)]^n \times H^s(\Omega)$. We then denote by $\mathscr{S}_n$ the operator (1.1) applying $(\vec{u}, p)$ on $(\vec{f}, g)$, and we write especially $\mathscr{S}_n[s, \Omega]$ for $\mathscr{S}_n$ acting from $D_n^s(\Omega)$ to $E_n^s(\Omega)$. We suppose everywhere that $s \neq \frac{1}{2}$.

3. General Fredholm properties. General Fredholm properties rely on general statements of [9] that we apply here to the Stokes system (1.1)-(1.2).

In [9], we develop general conditions for a strongly elliptic operator to be Fredholm between Sobolev spaces H^s (in the above sense; § 2.3), e.g., with Dirichlet conditions. Moreover, we extend that theory to strongly elliptic systems, and other ones satisfying a weaker ellipticity property that holds in particular for the Stokes system (see (7.7) in [9]).

We will apply those results. To do so, we recall the characteristic conditions concerning the operator and the domain. When Ω is a polygon, it is well known that such conditions are related to the angle openings of Ω and to associated discriminant functions (cf. [10], [12]). In fact it is related with the spectrum of a holomorphic operator family; in three dimensions the condition may be written only in that form. We show in [9] that those "spectral" conditions are fully convenient for "homogeneous" weighted Sobolev spaces, and that, for ordinary Sobolev spaces, they must be replaced with a new type of condition we call "injectivity modulo polynomials."

Although that distinction is of lesser use for regularity properties than for Fredholm properties, we introduce it in anticipation of the forthcoming Part II of this paper where we will describe the singularities of solutions.

Our conditions are related to tangent (or frozen) operators at each singular point of Ω.

3.1. Frozen operators at a vertex. Let Ω be a domain in $\mathscr{O}_n(\mathbf{R}^n)$, $n = 2, 3$ and $x \in A_0(\Omega)$. We will suppose that the diffeomorphism χ, which implies a neighborhood of x in Ω on a neighborhood of zero in Γ_x, is such that

$$D\chi(x) = I \text{ is the identity matrix.}$$

Then, the operator L_x, obtained by taking the principal part of the operator $\chi \circ \mathscr{S}_n \circ \chi^{-1}$ frozen in zero, just coincides with $\mathscr{S}_n$ on the cone Γ_x.

3.2. Frozen operators along an edge. As in the case of a vertex, if $x \in A_1(\Omega)$, the frozen operator on the wedge $\mathbf{R} \times \Gamma_x$ is $\mathscr{S}_3$. But, we have to define a new frozen operator L_x on the plane sector Γ_x (cf. [9, (3.3)]). Let (y, z) be coordinates such that $y \in \mathbf{R}$ and $z \in \Gamma_x$. The operator L_x is defined as

$$L_x(D_z) = \mathscr{S}_3(0, D_z)$$

(we remove tangential derivatives along the edge). Thus, we have

$$L_x(u_1, u_2, u_3, p) = (f_1, f_2, f_3, g) \tag{3.1}$$

if and only if

$$\mathscr{S}_2(u_1, u_2, p) = (f_1, f_2, g) \quad \text{and} \quad \Delta u_3 = f_3. \tag{3.2}$$

3.3. Injectivity modulo polynomials. For $\lambda \in \mathbf{C}$, $S_n^\lambda(\Gamma_x)$ denotes the set of vector functions $(u_1, \cdots, u_n, p)$ of the form:

$$u_j = r^\lambda \sum_{0 \leqq q \leqq Q} u_{jq}(\Psi) \log^q r \quad \text{with } u_{jq} \in \mathring{H}^1(G_x),$$

$$p = r^{\lambda - 1} \sum_{0 \leqq q \leqq Q} p_q(\Psi) \log^q r \quad \text{with } p_q \in L^2(G_x)$$

where $(r, \Psi) = (|z|, z/|z|)$ are the polar coordinates and G_x is the intersection of Γ_x with the unit sphere S^{n-1}.

We say that L is injective modulo polynomial on $S_n^\lambda(\Gamma_x)$ if

$$(\vec{u}, p) \in S_n^\lambda(\Gamma_x) \text{ and } L(\vec{u}, p) \text{ is polynomial implies that } (\vec{u}, p) \text{ is polynomial.}$$

Here "polynomial" means polynomial with respect to *cartesian variable* $z = (z_1, z_2)$ or (z_1, z_2, z_3). For instance, $r^\alpha \sin \alpha\theta$ is polynomial in $\mathbf{R}^2$ for $\alpha \in \mathbf{Z}$. Of course, the zero function is polynomial.

3.4. Index and regularity results.

THEOREM 3.3. *Let $\Omega \in \mathcal{O}_n(\mathbf{R}^n)$. The Stokes operator $\mathcal{S}_n[s, \Omega]$ is a Fredholm operator if and only if both the following conditions are satisfied*:

(3.4) $\forall x \in A_0(\Omega)$, $\forall \lambda$ *with* $\operatorname{Re} \lambda = s + 1 - n/2$, L_x *is injective modulo polynomials on* $S_n^\lambda(\Gamma_x)$;

(3.5) $\exists \varepsilon > 0$, $\forall x \in A_1(\Omega)$, $\forall \lambda$ *with* $\operatorname{Re} \lambda \in [0, s + \varepsilon]$, L_x *is injective modulo polynomials on* $S_3^\lambda(\Gamma_x)$.

This statement is derived from (7.15) in [9], with the variant (6.8) in [9].

If Ω has only conical points (which is the case when $n = 2$), the condition (3.5) is void. If Ω has no vertex (cf. examples (1.11), (1.12)), the condition (3.4) is void and (3.5) may be replaced with (3.5′):

(3.5′) $\forall x \in A_1(\Omega)$, $\forall \lambda$ with $\operatorname{Re} \lambda \in [0, s]$, L_x is injective modulo polynomials on $S_3^\lambda(\Gamma_x)$.

If Ω is a three-dimensional polyhedron with plane faces, (3.5) may still be replaced with (3.5′): the ε in (3.5) is useful in the case when Ω is a three-dimensional domain with smooth curved faces; that ε allows an easier formulation without introducing "subsections" or "singular chains," which describe the limit geometrical behavior at the neighborhood of a vertex.

THEOREM 3.6. *Assume that the conditions* (3.5) *and* (3.7) *are fulfilled*:

(3.7) $\forall x \in A_0(\Omega)$, $\forall \lambda$ *with* $\operatorname{Re} \lambda \in [1 - n/2, s + 1 - n/2]$, L_x *is injective modulo polynomials on* $S_n^\lambda(\Gamma_x)$.

Then, each solution $(\vec{u}, p) \in D_n^0(\Omega)$ of (1.1) *with $(\vec{f}, g)$ in $E_n^s(\Omega)$ has the regularity $D_n^s(\Omega)$.*

When (3.4) is satisfied, and not (3.7), there are singular functions. We will study these in Part II of this paper, along with the nonlinear Navier-Stokes system.

Now, we will study (3.5) and (3.7) in order to give more precise regularity results in two and three dimensions.

4. The link between the injectivity condition and the usual spectral condition.

4.1. Generalities. Let us study condition (3.4). In view of § 3.1, $L_x = \mathcal{S}_n$. If we consider $\vec{u}$ of the form $r^\lambda \vec{u}(\Psi)$ and $p = r^{\lambda-1} \not{p}(\Psi)$, then we get

$$\mathcal{S}_n(\vec{u}, p) = (\vec{f}, g)$$

where $\vec{f} = r^{\lambda-2} \vec{\not{f}}(\Psi)$ and $g = r^{\lambda-1} \not{g}(\Psi)$, with

$$\mathcal{L}_n(\lambda)(\vec{u}, \not{p}) = (\vec{\not{f}}, \not{g}), \tag{4.1}$$

$\mathcal{L}_n(\lambda)$ being a system on the sphere S^{n-1}, depending in a polynomial way on λ. As in [17], we can derive from the writing of $\mathcal{S}_n$ in polar coordinates that (4.1) may be written in the form

$$[\delta_n - \lambda(\lambda + 1)]\vec{u} + [(\lambda - 1)\vec{\Psi} + \vec{\nabla}_s]\not{p} = \vec{\not{f}}, \qquad \langle \lambda \vec{\Psi} + \vec{\sigma}, \vec{u} \rangle = \not{g}$$

where δ_n is the positive Laplace-Beltrami operator on S^{n-1}, $\vec{\Psi}$ is the vector $x/|x|$ in $\mathbf{R}^n$, and $\vec{\nabla}_s$ is the tangential component of the gradient on the sphere

$$\vec{\nabla}_s = \vec{\nabla} - \vec{\Psi}\partial/\partial_r.$$

$\mathscr{L}_n(\lambda)$ gives rise to an operator acting from $D_n^0(G_x)$ to $E_n^0(G_x)$. It is almost everywhere invertible. The set of λ for which $\mathscr{L}_n(\lambda)$ is not invertible is called the **spectrum** of $\mathscr{L}_n$, and the condition used by Kondrat'ev [12] or Maz'ja and Plamenevskii is that the straight line $\operatorname{Re}\lambda = s+1-n/2$ does not meet the spectrum of $\mathscr{L}_n$. As we have already said, this type of condition is correct for weighted Sobolev spaces (of the type $r^{\gamma+|\alpha|}D^\alpha v \in L^2$), but it is not always suitable for ordinary Sobolev spaces. Nevertheless, we have (cf. (4.2) for $s=0$ and (4.6) in [9]):

LEMMA 4.2. *If λ is not a positive integer, $\mathscr{S}_n$ is injective modulo polynomials on $S^\lambda(\Gamma_x)$ if and only if λ does not belong to the spectrum of $\mathscr{L}_n$ on G_x.*

If λ is an integer number, the comparison depends on the difference $d(\lambda)$ between the dimensions of two spaces of polynomial functions:

$$d(\lambda) = \dim P^\lambda(\Gamma_x) - \dim Q^{\lambda-2}$$

where $P^\lambda(\Gamma_x)$ is the set of the elements of $S^\lambda(\Gamma_x)$ that are polynomials *in cartesian variables*, and $Q^{\lambda-2}$ is the set of the $(\vec{f}, g)$ with f_j (respectively, g) homogeneous polynomial of degree $\lambda-2$ (respectively, $\lambda-1$) in z. $d(\lambda)$ depends only on Γ_x. According to [9, Annex D], there exists a homogeneous polynomial A that is zero on the boundary of Γ_x and such that if B is a polynomial that is zero on $\partial\Gamma_x$, then A divides B (i.e., $P^\lambda(\Gamma_x)$ is a principal ideal).

If the degree of A is two, then $d(\lambda)=0$; and according to (4.9) and (7.14) in [9], we have the following lemma.

LEMMA 4.3. *If $d^0A=2$, for each integer λ we have the same equivalence as in* (4.2).

According to (4.8) and (7.14) in [9], we have the following lemma.

LEMMA 4.4. *If $d^0A \geqq 3$, for $\lambda=1$ we have the same equivalence as in* (4.2); *but for each integer $\lambda \geqq 2$, $\mathscr{S}_n$ is not injective modulo polynomials on $S^\lambda(\Gamma_x)$.*

According to (4.10) and (7.14) in [9], we have Lemma 4.5.

LEMMA 4.5. *If $d^0A=1$, and $\lambda \in \mathbf{N}^*$, then $\mathscr{S}_n$ is injective modulo polynomials on $S^\lambda(\Gamma_x)$ if and only if $\mathscr{L}_n(\mu)^{-1}$ has a pole of order one in $\mu=\lambda$ and if*

$$\dim \operatorname{Ker} \mathscr{L}_n(\lambda) = d(\lambda).$$

4.2. Application to two-dimensional cones. It is well known that the poles $\mathscr{L}_2(\lambda)^{-1}$ coincide with the roots of the following equations:

$$F_\omega(\lambda)=0 \quad \text{where} \quad F_\omega(\lambda) = \frac{\sin^2\lambda\omega - \lambda^2\sin^2\omega}{\lambda^2} \tag{4.6}$$

because, more precisely, $\mathscr{L}_2(\lambda)^{-1}F_\omega(\lambda)$ is holomorphic on $\mathbf{C}$ (cf. [12], [10], [11], [5]).

When the opening ω of the plane sector Γ_x is not 2π, then the two sides of Γ_x are independent and $d^0A=2$. So, the condition (3.4) is that F_ω has no zero with real part s.

When $\omega=2\pi$, then $d^0A=1$ and $d(\lambda)>0$ for all positive integer numbers. As in § 15.B in [9] for fourth-order operators, we show in the Annex that, according to Lemma 4.5, $\mathscr{S}_2$ is injective modulo polynomials on S^λ for each $\lambda \in \mathbf{N}^*$ (including $\lambda=1$). As the roots of (4.6) are the half integers, we find that condition (3.4) is reduced to

$$s \neq k+\tfrac{1}{2}, \quad \forall k \in \mathbf{N}.$$

4.3. Application to three-dimensional cones. If Γ_x is a revolution cone, then $d^0A=2$ and we apply Lemmas 4.2 and 4.3.

If Γ_x is a polyhedral cone, let D be the number of distinct planes containing at least one side of Γ_x. For a cube, $D=3$. For a pyramid with a square basis, $D=4$. If $D \geqq 3$, we apply Lemmas 4.2 and 4.4.

5. Precise results in two-dimensional domains.

5.1. More about the discriminant function F_ω. The roots of (4.6) have been studied by Seif [21], Lozi [13], Dauge [7], Bernardi and Raugel [2], [3], and Maslovskaya [14]. Bernardi and Raugel give a table for the roots of (4.6) with the lowest positive real part. Here, we complete that work by a table for the roots of (4.6) with their real parts $\xi \in [0, 4]$ and by the corresponding graph (Fig. 2) of ξ in function of ω.

Let us denote λ by $\xi + i\eta$, with ξ, η real. We are interested in roots of (4.6) with $\xi \geqq 0$. $\lambda = 1$ is always a root of (4.6) and plays a particular role (see § 5.2).

We denote by $\xi_k(\omega)$ the real part of the kth root of

$$(\lambda - 1)^{-1} F_\omega(\lambda) = 0$$

the roots being ordered with increasing real part and repeated according to their multiplicities. The following can be shown (cf. [7], [2]):

(a) If $\omega \in]0, \pi[$, $\xi_1(\omega) > \pi/\omega$.

(b) If $\omega \in]\pi, 2\pi[$, $\xi_1(\omega) \in]\sup(\frac{1}{2}, \omega_1/\omega), \pi/\omega[$, where $\omega_1 = 0.812825\pi$; ω_1 is the root of

$$\omega \in]0, \pi[\quad \text{and} \quad \frac{\sin \omega}{\omega} = -\cos \omega_0 \quad \text{with } \tan \omega_0 = \omega_0.$$

Tables 1 and 2 and Fig. 2 give values for $\xi_1, \cdots, \xi_{14}$ that occur in $[0, 4]$. A dash means a value greater than four.

For $j \geqq 1$, let I_j be the set of ω such that $\xi_{2j}(\omega)$ and $\xi_{2j+1}(\omega)$ coincide. In the interior of I_j, ξ_{2j} and ξ_{2j+1} are the real parts of two conjugate nonreal numbers. For $\omega \in \partial I_j$ and $\omega \neq 0$, there is a real double root and the bifurcation of two real roots.

TABLE 1

ω	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6
0.4	3.397	3.397	-	-	-	-
0.5	2.740	2.740	4.808	4.808	-	-
0.6	2.307	2.307	4.022	4.022	-	-
0.7	2.004	2.004	3.464	3.464	-	-
0.8	1.783	1.783	3.051	3.051	4.312	4.312
0.9	1.252	1.988	2.542	2.932	3.853	3.853
1	1	2	2	3	3	4
1.1	0.834	1.662	2.012	2.475	3.096	3.215
1.2	0.718	1.408	2.045	2.045	2.883	2.883
1.3	0.637	1.207	1.882	1.882	2.657	2.657
1.4	0.581	1.044	1.745	1.745	2.465	2.465
1.5	0.544	0.909	1.629	1.629	2.301	2.301
1.6	0.522	0.796	1.530	1.530	2.159	2.159
1.7	0.509	0.701	1.444	1.444	2.035	2.035
1.8	0.503	0.622	1.258	1.480	1.927	1.927
1.9	0.500	0.555	1.111	1.498	1.670	1.994
2	0.5	0.5	1	1.5	1.5	2

TABLE 2

ω	ξ_7	ξ_8	ξ_9	ξ_{10}	ξ_{11}	ξ_{12}
1.1	4.067	4.067	-	-	-	-
1.2	3.720	3.720	-	-	-	-
1.3	3.430	3.430	4.203	4.203	-	-
1.4	3.184	3.184	3.901	3.901	-	-
1.5	2.972	2.972	3.641	3.641	-	-
1.6	2.788	2.788	3.415	3.415	4.042	4.042
1.7	2.626	2.626	3.216	3.216	3.806	3.806
1.8	2.484	2.484	3.041	3.041	3.597	3.597
1.9	2.233	2.485	2.808	2.964	3.413	3.413
2	2	2.5	2.5	3	3	3.5

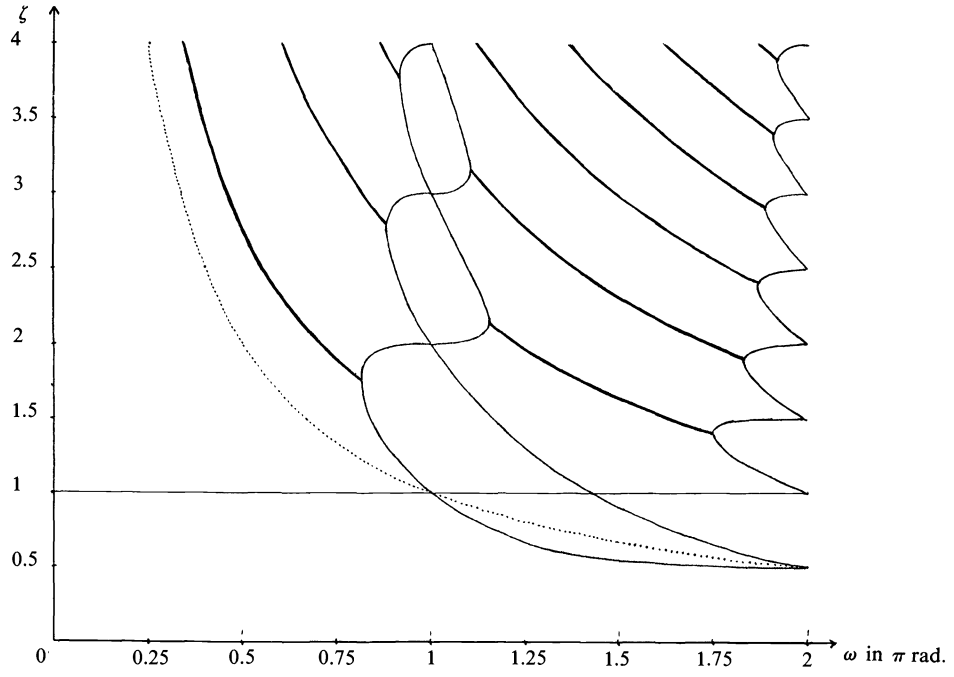


FIG. 2

I_1 has only one connected component: $I_1 =]0, \omega_1]$. For $j \geqq 2$, I_j has two connected components: $]0, \omega_j]$ and $[\omega_j', \omega_j'']$. When $j \to +\infty$, $\omega_j \to \pi$ and $\omega_j'' \to 2\pi$ increase, while $\omega_j' \to \pi$ decreases. All integers are double roots for $\omega = \pi$, and all half integers are double roots for $\omega = 2\pi$.

On the graph, the dotted line is the graph of $\omega \to \pi/\omega$. The heavy lines represent a double value for the ξ_k (conjugate roots), and the ordinary lines represent real roots.

Table 3 gives the values of the ω_j, ω_j', ω_j'' that occur in Fig. 2.

5.2. The special case of the pole $\lambda = 1$. As we have already shown $\lambda = 1$ is always a pole for $\mathscr{L}_2(\lambda)^{-1}$. But, if the opening of the cone Γ is $\omega = 2\pi$, $\mathscr{S}_2$ is injective modulo polynomials on $S^\lambda(\Gamma)$. If $\omega \neq 2\pi$, this is not so for $\mathscr{S}_2$.

It is easy to show that $\operatorname{Ker} \mathscr{L}_2(1)$ is one-dimensional and is generated by $(\vec{0}, 1)$ (see [11], [5]). As a consequence, we get the following lemma.

TABLE 3

	ω	ω'	ω''
1	0.813	-	-
2	0.884	1.154	1.751
3	0.915	1.102	1.825

LEMMA 5.1. *Let s be such that $1 \leqq s < \xi_1(\omega)$. Let $(\vec{u}, p)$ be in $D_2^0(\Gamma)$ such that $\mathcal{S}_2(\vec{u}, p) = (\vec{f}, g) \in E_2^s(\Gamma)$. Moreover, if $s > 1$, we suppose that $g(0) = 0$; if $s = 1$ we suppose that $r^{-1}g \in L^2(\Gamma)$. Then, if B denotes the unit ball, we have*

$$(\vec{u}, p) \in D_2^s(\Gamma \cap B).$$

Proof. We derive the proof from the methods of [9]. For $s = 1$, it is the result of [11].

By a cut-off, we assume that $(\vec{u}, p)$ has compact support. We use the Mellin transform $\mathcal{M}$ of $(\vec{u}, rp)$, which is defined for $\operatorname{Re} \lambda \leqq 0$; we have

$$\mathcal{L}_2(r\partial_r)(\vec{u}, rp) = (r^2\vec{f}, rg), \tag{5.2}$$

and thus using the Mellin transform we have

$$\mathcal{L}_2(\lambda)(\vec{U}(\lambda), P(\lambda)) = (\vec{F}(\lambda), G(\lambda)). \tag{5.3}$$

But $(\vec{F}(\lambda), G(\lambda))$ is defined for $\operatorname{Re} \lambda < s$. If $s \neq 1$, then we deduce from [9] (see the condensed proof in [8]) that there exists $(\vec{u}_0, p_0) \in D_2^s(\Gamma \cap B)$ such that for $\operatorname{Re} \lambda = 1 + \varepsilon$ with $\varepsilon \in]0, \min(1, s-1)[$,

$$\mathcal{M}(\vec{u}_0, rp_0)(\lambda) = (\vec{U}(\lambda), P(\lambda))$$

where $(\vec{U}, P)(\lambda)$ is the extension determined by (5.3). And we have

$$(\vec{u}, p) - (\vec{u}_0, p_0) = \sum_{0 < \operatorname{Re}\lambda < 1+\varepsilon} \operatorname{Res}\,[r^\lambda \vec{U}(\lambda), r^{\lambda-1}P(\lambda)].$$

Since $s < \xi_1(\omega)$, the sum is reduced to $\lambda = 1$. And we have

$$\begin{aligned} \mathcal{S}_2 \operatorname{Res}\,(r^\lambda \vec{U}(\lambda), r^{\lambda-1}P(\lambda)) &= \operatorname{Res}\,\mathcal{S}_2(r^\lambda \vec{U}(\lambda), r^{\lambda-1}P(\lambda)) \\ &= \operatorname{Res}_{\lambda=1}\,(r^{\lambda-2}\vec{F}(\lambda), r^{\lambda-1}G(\lambda)) \\ &= (\vec{0}, g(0)). \end{aligned}$$

The second equality is given by (5.3) and by the equivalence of $\mathcal{S}_2(\vec{u}, p) = (\vec{f}, g)$ with (5.2). Since $g(0) = 0$, we get

$$\mathcal{S}_2 \operatorname{Res}_{\lambda=1}\,(r^\lambda \vec{U}(\lambda), r^{\lambda-1}P(\lambda)) = 0.$$

Therefore

$$\mathcal{L}_2(1) \operatorname{Res}_{\lambda=1}\,(\vec{U}(\lambda), P(\lambda)) = 0.$$

The residue belongs to the kernel of $\mathcal{L}_2(1)$. Thus it is equal to $(\vec{0}, c)$, with c a constant. We finally get

$$(\vec{u}, p) - (\vec{u}_0, p_0) = (\vec{0}, c).$$

Thus $(\vec{u}, p) \in D_2^s(\Gamma \cap B)$.

If $s = 1$, since $r^{-1}g \in L^2$, $G(\lambda)$ is defined up to $\operatorname{Re} \lambda = 1$ (such is not the case if $g \in H^1$ only). Since $\operatorname{Res}_{\lambda=1} \mathcal{L}_2(\lambda)^{-1} = (\vec{0}, 1)\chi$, where χ is a linear form, we get that $\Pi\mathcal{L}_2(\lambda)^{-1}$ is holomorphic in the neighborhood of $\lambda = 1$, where Π is the projection on

the speed component. Thus, $\Pi\mathscr{L}_2(\lambda)^{-1}(F, G)(\lambda)$ is defined up to $\operatorname{Re}\lambda = 1$, with suitable estimates. Then, we deduce that $\vec{u} \in H^2(\Gamma \cap B)$. Since $p \in L^2$, and $\vec{\nabla} p = \Delta\vec{u} + \vec{f} \in L^2(\Gamma \cap B)$, then $p \in H^1(\Gamma \cap B)$. □

5.3. Index and regularity results.

THEOREM 5.4. *Let $\Omega \in \mathscr{O}_2(\mathbf{R}^2)$, and let $s > 0$. $\mathscr{S}_2(\Omega, s)$ is Fredholm if and only if the three following conditions are fulfilled*:

(a) $s \neq 1$;

(b) $\forall x \in A_0(\Omega)$ *such that* $\omega_x \neq 2\pi$, $\forall k$, $s \neq \xi_k(\omega_x)$;

(c) $\forall x \in A_0(\Omega)$ *such that* $\omega_x = 2\pi$, $\forall k$, $s \neq k + \frac{1}{2}$.

Let us recall that ξ_k is defined in § 5.1. It is a straightforward consequence of Theorem 3.3 and §§ 4.2, 4.3, and 5.1. From Theorem 3.6 and § 5.2 we derive Theorem 5.5.

THEOREM 5.5. *Let $\Omega \in \mathscr{O}_2(\mathbf{R}^2)$ and $s > 0$. Let $(\vec{u}, p) \in D_2^0(\Omega)$ be the solution of* (1.1) *with $(\vec{f}, g) \in E_2^s(\Omega)$*:

(a) **If** $s < 1$ *and* $s < \min_{x \in A_0(\Omega)} \xi_1(\omega_x)$, *then* $(\vec{u}, p) \in D_2^s(\Omega)$;

(b) **If** $s > 1$ *and moreover* $g(x) = 0$ *for each vertex* x, *and if* $s < \min_{x \in A_0(\Omega)} \xi_1(\omega_x)$, *then* $(\vec{u}, p) \in D_2^s(\Omega)$;

(c) **If** $s = 1$ *and moreover* $r_x^{-1} g \in L^2(\Omega)$ *for each vertex* x, *and if* $1 < \min_{x \in A_0(\Omega)} \xi_1(\omega_x)$, *then* $(\vec{u}, p) \in D_2^2(\Omega)$.

As $\xi_1(\pi) = 1$ and ξ_1 is a decreasing function, $1 < \min_{x \in A_0(\Omega)} \xi_1(\omega_x)$ holds if Ω is convex. It coincides with the result in [11].

6. Precise results in three-dimensional domains when there are edges, but no vertex.

6.1. The statements. In such a case, we study condition (3.5): since, for $x \in A_1(\Omega)$, L_x is given by (3.1)-(3.2), it is obvious that L_x is injective modulo polynomials on $S_3^\lambda(\Gamma_x)$ and only if we have (6.1) and (6.2):

(6.1) $\mathscr{S}_2$ is injective modulo polynomials on $S_2^\lambda(\Gamma_x)$,

(6.2) Δ is injective modulo polynomials on $S_0^\lambda(\Gamma_x)$,

where $S_0^\lambda(\Gamma) = \{v = r^\lambda \sum v_q(\Psi) \log^q r,\ v_q \in \mathring{H}^1(G)\}$.

If $\omega_x \neq 2\pi$, (6.2) is equivalent to $\lambda \notin \{(k\pi/\omega_x)/k \in \mathbf{N}^*\}$ and if $\omega_x = 2\pi$, (6.2) is equivalent to $\lambda \notin \{k + \frac{1}{2},\ k \in \mathbf{N}\}$ (for $\operatorname{Re}\lambda \geqq 0$). Our statement follows.

THEOREM 6.3. *Let $\Omega \in \mathscr{O}_3(\mathbf{R}^3)$ such that $A_0(\Omega) = \varnothing$. Let $(\vec{u}, p) \in D_3^0(\Omega)$ be the solution of* (1.1) *with $(\vec{f}, g) \in E_3^s(\Omega)$.*

(a) **If** $s < 1$ *and* $s < \min_{x \in A_1(\Omega)} \xi_1(\omega_x)$, *then* $(\vec{u}, p) \in D_3^s(\Omega)$;

(b) **If** $s > 1$ *and* $s < \min_{x \in A_1(\Omega)} \pi/\omega_x$, *and moreover* $g(x) = 0$ *for each* $x \in A_1(\Omega)$, *then* $(\vec{u}, p) \in D_3^s(\Omega)$;

(c) **If** $s = 1$ *and* Ω *is convex, and moreover* $\rho_1^{-1} g \in L^2(\Omega)$, *where* ρ_1 *is the distance from* $A_1(\Omega)$, *then* $(\vec{u}, p) \in D_3^2(\Omega)$.

Proof. First, here are the main arguments of the proof.

(a) If $s < 1$, then (6.1) (respectively, (6.2)) is true for all λ such that $\operatorname{Re}\lambda \in [0, s]$ if $s < \xi_1(\omega_x)$ (respectively, $s < \pi/\omega_x$). But, if $\xi_1(\omega_x) < 1$, then $\xi_1(\omega_x) < \pi/\omega_x$. Thus (a) is derived from (3.6).

(b) As $s > 1$, if $s < \pi/\omega_x$, then $\omega_x < \pi$ and $\xi_1(\omega_x) > \pi/\omega_x$. Thus, for each $\lambda \neq 1$ in the strip $\operatorname{Re}\lambda \in [0, s]$, and for each $x \in A_1(\Omega)$, L_x is injective modulo polynomials on $S_3^\lambda(\Gamma_x)$. We have that (b) is an adaptation of the proof of (5.12) in [9] by taking advantage of Lemma 5.1 above. See some details that follow this proof.

(c) When Ω is convex, $\xi_1(\omega_x)$ and π/ω_x are greater than one for each vertex x. Like (b), (c) is derived from Lemma 5.1.

6.2. A more detailed proof of Theorem 6.3. Because of the special role of the pole $\lambda = 1$ (cf. Lemma 5.1) we are led to revise the proofs in § 12.C of [9] to take into account the cancellation assumptions concerning g.

Let Γ be a plane sector. We first study $\mathcal{S}_3$ on $\mathbf{R}\times\Gamma$. We denote by y the coordinate in $\mathbf{R}$ and by z the coordinates in Γ; $r = |z|$. We introduce the space $H_0^1(\mathbf{R}\times\Gamma)$ as the set of functions v such that

$$\forall \alpha, |\alpha| \leqq 1 \quad r^{|\alpha|-1} D^\alpha v \in L^2(\mathbf{R}\times\Gamma).$$

$\mathring{H}_0^1(\mathbf{R}\times\Gamma)$ is the closure of $\mathcal{D}(\mathbf{R}\times\Gamma)$ in $H_0^1(\mathbf{R}\times\Gamma)$ and $H_0^{-1}(\mathbf{R}\times\Gamma)$ is its dual space.

LEMMA 6.4. *$\mathcal{S}_3$ induces an isomorphism from $\mathring{H}_0^1\times\mathring{H}_0^1\times\mathring{H}_0^1\times L^2(\mathbf{R}\times\Gamma)$ to $H_0^{-1}\times H_0^{-1}\times H_0^{-1}\times L^2(\mathbf{R}\times\Gamma)$.*

Sketch of the Proof (cf. [9, (8.1), (12.6)]). It is sufficient to prove that $\mathcal{S}_3$ is injective and has a closed range because its adjoint has the same form. Let B_ρ be the ball with center zero and radius ρ. For $(\vec{u}, p) \in (\mathring{H}^1)^3 \times L^2$ with its support in B_1, we have the following estimate:

$$\sum_{j=1}^{3} \|u_j\|_{H^1} + \|p\|_{L^2} \leqq C\left(\sum_{j=1}^{3} \|f_j\|_{H^{-1}} + \|g\|_{L^2} + \sum_{j=1}^{3} \|u_j\|_{L^2} + \|p\|_{H^{-1}} \right).$$

But $\mathring{H}^1(B_1) \subset H_0^1(B_1)$ and $H_0^{-1}(B_1) \subset H^{-1}(B_1)$. Moreover, if $\operatorname{supp}(\vec{u}, p) \subset B_\rho$, we have

$$\|u_j\|_{L^2} \leqq \rho \|u_j\|_{H_0^1} \quad \text{and by duality} \quad \|p\|_{H^{-1}} \leqq \rho \|p\|_{L^2}.$$

Thus, if $\operatorname{supp}(\vec{u}, p) \subset B_\rho$, we have the following estimate, for $\rho > 0$ small enough:

$$\sum \|u_j\|_{H_0^1} + \|p\|_{L^2} \leqq 2C\left(\sum \|f_j\|_{H_0^{-1}} + \|g\|_{L^2} \right). \tag{6.5}$$

We deduce the same estimates for all $(\vec{u}, p)$ in $(\mathring{H}_0^1)^3 \times L^2$ by homogeneity and density of functions with compact support. □

$\mathcal{S}_3$ may be written $\mathcal{S}_3(D_y, D_z)$.

LEMMA 6.6. *$\mathcal{S}_3(\pm 1, D_z)$ induces an isomorphism from $\mathring{H}^1\times\mathring{H}^1\times\mathring{H}^1\times L^2(\Gamma)$ to $H^{-1}\times H^{-1}\times H^{-1}\times L^2(\Gamma)$.*

Sketch of the Proof. Let $(\vec{u}, p) \in (\mathring{H}^1)^3 \times L^2(\Gamma)$. With φ a cut-off function in $\mathbf{R}$, $\varphi \equiv 1$ in the neighborhood of zero, we consider

$$(\vec{v}, q)(y, z) = \varphi(y)\, e^{i\rho y} (\vec{u}, p)(z),$$

for $\rho \geqq 1$ and apply the estimate (6.5) to $(\vec{v}, q)$. Denoting $(\vec{f}, g) = \mathcal{S}_3(\rho, D_z)(\vec{u}, p)$, we have

$$\mathcal{S}_3(\vec{v}, q) = \varphi(y)\, e^{i\rho y} (\vec{f}, g)(z) + e^{i\rho y} (\vec{f}_R, g_R).$$

Let us introduce the $H_0^1(\Gamma, \rho)$-norm on $\mathring{H}^1(\Gamma)$:

$$\|W\|^2_{H_0^1(\Gamma,\rho)} \equiv \rho^2 \|W\|^2_{L^2} + \|r^{-1} W\|^2_{L^2} + \sum_{|\alpha|=1} \|D^\alpha W\|^2_{L^2} \tag{6.7}$$

and $H_0^{-1}(\Gamma, \rho)$ the dual norm. We have the equivalences:

$$\|v_j\|_{H_0^1(\mathbf{R}\times\Gamma)} \approx \|u_j\|_{H_0^1(\Gamma,\rho)} \quad \text{and} \quad \|\varphi\, e^{i\rho y} f_j\|_{H_0^{-1}(\mathbf{R}\times\Gamma)} \approx \|f_j\|_{H_0^{-1}(\Gamma,\rho)}.$$

So the estimate (6.5) implies that

$$\begin{aligned} \sum \|u_j\|_{H_0^1(\Gamma,\rho)} + \|p\|_{L^2(\Gamma)} \leqq C\Big(& \sum \|f_j\|_{H_0^{-1}(\Gamma,\rho)} + \|e^{i\rho y} f_{R,j}\|_{H_0^{-1}(\mathbf{R}\times\Gamma)} \\ & + \|g\|_{L^2(\Gamma)} + \|e^{i\rho y} g_R\|_{L^2(\mathbf{R}\times\Gamma)} \Big). \end{aligned} \tag{6.8}$$

The support of g_R is included in supp φ, and

$$g_R(y, z) = \sum a_j(y) u_j(z), \text{ with } a_j \text{ smooth.}$$

So,

$$\| e^{i\rho y} g_R \|_{L^2(\mathbf{R}\times\Gamma)} \leqq C\rho^{-1} (\sum \| u_j \|_{H^1_0(\Gamma,\rho)} + \| p \|_{L^2(\Gamma)}),$$

and we can prove the same estimate for $\| e^{i\rho y} f_{R,j} \|$ in (6.8). So, for ρ large enough,

$$\sum \| u_j \|_{H^1_0(\Gamma,\rho)} + \| p \|_{L^2(\Gamma)} \leqq 2C (\sum \| f_j \|_{H_0^{-1}(\Gamma,\rho)} + \| g \|_{L^2(\Gamma)}),$$

which is an a priori estimate for $\mathcal{S}_3(\rho; D_z)$. Using a suitable scaling, we get an estimate for $\mathcal{S}_3(1, D_z)$. The proof for $\mathcal{S}_3(-1, D_z)$ is the same, with $e^{-i\rho y}$ instead of $e^{i\rho y}$. □

For $s > 1$, we replace $E^s_3(\Gamma)$ by $F^s_3(\Gamma)$, which is the space of the $(\vec{f}, g) \in E^s_3(\Gamma)$ such that $g(0) = 0$. For $s = 1$, $F^1_3(\Gamma)$ is characterized by $r^{-1} g \in L^2(\Gamma)$. When $\mathcal{S}_3(1, D_z)$-$(\vec{u}, p) = (\vec{f}, g)$, we have

$$\partial_1 u_1 + \partial_2 u_2 + i u_3 = g.$$

Let us suppose that the opening ω of Γ is not 2π. When the u_j belong to $H^{s+1}(\Gamma)$, with $s > 1$, since they are zero on $\partial\Gamma$, $\nabla u_j(0) = 0$; thus $g(0) = 0$. On the other hand, if $s = 1$ and $u_j \in H^2 \cap \mathring{H}^1(\Gamma)$, then $r^{|\alpha|-2} D^\alpha u_j \in L^2(\Gamma)$ for $|\alpha| \leqq 2$ (cf. [9 (AC.6)]; thus $r^{-1} g \in L^2(\Gamma)$.

Therefore $\mathcal{S}_3(\pm 1, D_z)$ operate from $D^s_3(\Gamma)$ to $F^s_3(\Gamma)$. As a consequence of Lemmas 6.6 and 5.1, we get Proposition 6.9.

PROPOSITION 6.9. *Let $s \geqq 1$ be such that $s < \pi/\omega$. Then $\mathcal{S}_3(\pm 1, D_z)$ is an isomorphism from $D^s_3(\Gamma)$ to $F^s_3(\Gamma)$.*

By partial Fourier transform with respect to y the equation $\mathcal{S}_3(D_y, D_z)(u, p) = (f, g)$ becomes

$$\mathcal{S}_3(\pm\rho, D_z)(\hat{u}, \hat{p})(\pm\rho, z) = (\hat{f}, \hat{g})(\pm\rho, z), \qquad \rho \geqq 0.$$

We define $F^s_3(\mathbf{R}\times\Gamma)$ by the condition $g(y, 0) = 0$ if $s > 1$ and $|z|^{-1} g(y, z) \in L^2(\mathbf{R}\times\Gamma)$ if $s = 1$. If $(f, g) \in F^s_3(\mathbf{R}\times\Gamma)$, then for all ρ, $(\hat{f}, \hat{g})(\pm\rho) \in F^s_3(\Gamma)$. By a scaling argument, we deduce from Proposition 6.9 the uniform estimate for $\rho \geqq 1$:

$$\| (\hat{u}, \hat{p})(\pm\rho) \|_{D^s_3(\Gamma,\rho)} \leqq C \| (\hat{f}, \hat{g})(\pm\rho) \|_{F^s_3(\Gamma,\rho)}$$

where $H^s(\Gamma, \rho)$ means the norm $\| \ \|_{H^s} + \rho^s \| \ \|_{L^2}$ which obviously defines $D^s_3(\Gamma, \rho)$ and $F^s_3(\Gamma, \rho)$ when $s \neq 1$; $F^1_3(\Gamma, \rho) = L^2(\Gamma)^3 \times H^1_0(\Gamma, \rho)$ (see (6.7)). We also have an a priori estimate for $\rho \leqq 1$.

So, in the same way as in 9.C of [9], we get the following lemma.

LEMMA 6.10. *$1 \leqq s < \pi/\omega$. $\Gamma_\rho \equiv \Gamma \cap B_\rho$. Then the inverse operator $(\mathcal{S}_3)^{-1}$ of (6.4) induces a continuous operator from $F^s_3(\mathbf{R}\times\Gamma_2)$ to $D^s_3(\mathbf{R}\times\Gamma_1)$.*

Now, if we go back to the operator $\mathcal{S}_3$ on Ω, for each $x \in A_1(\Omega)$, we get an operator equal to $\mathcal{S}_3 + \mathcal{T}$, $\mathcal{T}$ being a perturbation. It is important to note that the fourth equation of $(\mathcal{S}_3 + \mathcal{T})(\vec{u}, p) = (\vec{f}, g)$ has the following form:

$$\sum_{1 \leqq j,k \leqq 3} a_{j,k}(y, z) \partial_j u_k = g \quad \text{with } a_{j,k} \text{ smooth.}$$

Thus, if $(\vec{u}, p) \in D^s_3(\mathbf{R}\times\Gamma)$, then $(\vec{f}, g) \in F^s_3(\mathbf{R}\times\Gamma)$. So, we are able to use Lemma 6.10 along with the perturbation argument and Neumann series of 10.D in [9] in order to get the local regularity result in the neighborhood of each $x \in A_1(\Omega)$.

7. Study of the parametrical operator associated to the Stokes system in three-dimensional domains: Description of areas free of poles.

7.1. First identities. Let Γ be a cone in $\mathbf{R}^3$ and let G be its intersection with the unit sphere S^2.

In § 4.1 we introduced an operator $\mathscr{L}_3(\lambda)$ acting from $D_3^0(G)$ to $E_3^0(G)$. In view of Lemmas 4.2 and 4.3, we wish to find areas in $\mathbf{C}$ where $\mathscr{L}_3(\lambda)$ is everywhere invertible. As the index of $\mathscr{L}_3(\lambda)$ is zero, it is equivalent to find where $\mathscr{L}_3(\lambda)$ is injective. But, from the definition of $\mathscr{L}_3(\lambda)$ we deduce that

$$\mathscr{L}_3(\lambda)(\mathbf{u}, p) = 0 \Leftrightarrow \mathscr{S}_3(r^\lambda \mathbf{u}, r^{\lambda-1} p) = 0.$$

For $\operatorname{Re}\lambda = -\frac{1}{2}$, $\mathscr{L}_3(\lambda)$ is always injective; it is a consequence of Theorem 3.3 for $s = 0$ (see also [16]). On the other hand, according to condition (3.7), we are interested in the strip $\operatorname{Re}\lambda \in [-\frac{1}{2}, s - \frac{1}{2}]$. Thus, we suppose the following:

$$\operatorname{Re}\lambda > -\tfrac{1}{2}, \tag{7.1}$$

$$(\mathbf{u}, p) \in (\mathring{H}^1(G))^3 \times L^2(G), \qquad (\mathbf{u}, p) \neq 0, \tag{7.2}$$

$$-\Delta(r^\lambda \mathbf{u}) + \nabla(r^{\lambda-1} p) = 0, \tag{7.3}$$

$$\operatorname{div}(r^\lambda \mathbf{u}) = 0. \tag{7.4}$$

We denote

$\xi = \operatorname{Re}\lambda$, $\eta = \operatorname{Im}\lambda$;

z the cartesian coordinates in $\mathbf{R}^3$; $\Psi = z/|z|$;

$\mathbf{u}_r = \langle \vec{\mathbf{u}}, \vec{\Psi} \rangle$, the radial component of $\mathbf{u}$;

δ the positive Laplace-Beltrami operator on $\mathring{H}^1(G)$;

∇_s the spherical part of the gradient.

If C is $\Gamma \cap \{1 < r < 2\}$, we get by integrating (7.3) with $r^{\bar\lambda}\bar{\mathbf{u}}$ and (7.4) with $r^{\bar\lambda - 1}\bar p$:

$$\int_C \langle -\Delta(r^\lambda \mathbf{u}) + \nabla(r^{\lambda-1} p), r^{\bar\lambda}\bar{\mathbf{u}} \rangle + \operatorname{div} r^\lambda \mathbf{u} \cdot r^{\bar\lambda - 1}\bar p = 0.$$

As in [16], it implies by integration by parts:

$$\int_G |\nabla_s \mathbf{u}|^2 - \lambda(\lambda+1)|\mathbf{u}|^2 + (2\xi + 1)\int_G p\bar{\mathbf{u}}_r = 0. \tag{7.5}$$

And, again as in [16], we deduce from

$$\langle -\Delta(r^\lambda \mathbf{u}) + \nabla(r^{\lambda-1} p), z \rangle = 0$$

and from (7.4) that

$$\delta \mathbf{u}_r - (\lambda+1)(\lambda+2)\mathbf{u}_r + (\lambda - 1)p = 0. \tag{7.6}$$

By integrating (7.6) on G with $\bar{\mathbf{u}}_r$, we get

$$\int_G |\nabla_s \mathbf{u}_r|^2 - (\lambda+1)(\lambda+2)|\mathbf{u}_r|^2 + (\lambda - 1)p\bar{\mathbf{u}}_r = 0. \tag{7.7}$$

And from (7.4), which implies

$$\int_C \operatorname{div}(r^\lambda \mathbf{u}) = 0,$$

we get

$$(\lambda + 2)\int_G \mathbf{u}_r = 0. \tag{7.8}$$

7.2. The case when λ is not real.

LEMMA 7.9. *Let us assume that we have* (7.1)-(7.4) *and moreover that* $\eta \neq 0$. *Then*

$$\xi \geqq \frac{1}{\sqrt{3}}(\Lambda_1 + 1 + \eta^2)^{1/2}$$

where Λ_1 *is the first eigenvalue of* δ.

Proof. We take the imaginary part of (7.5) and divide it by $(2\xi+1)$ and obtain

$$\text{Im} \int_G p\bar{\mathbf{u}}_r = \eta \int_G |\mathbf{u}|^2. \tag{7.10}$$

We take the imaginary part of (7.7) and use (7.10); after we divide that by η we obtain

$$-(2\xi+3)\int_G |\mathbf{u}_r|^2 + (\xi-1)|\mathbf{u}|^2 + \text{Re}\int_G p\bar{\mathbf{u}}_r = 0. \tag{7.11}$$

We take the real part of (7.5) and eliminate $\text{Re}\int_G p\bar{\mathbf{u}}_r$ by using (7.11). Then we obtain

$$\int_G |\nabla_s \mathbf{u}|^2 + [\eta^2 - \xi(\xi+1) - (2\xi+1)(\xi-1)]|\mathbf{u}|^2 + (2\xi+1)(2\xi+3)\int_G |\mathbf{u}_r|^2 = 0.$$

Since $(2\xi+1)(2\xi+3) > 0$, it implies

$$\int_G |\nabla_s \mathbf{u}|^2 \leqq (3\xi^2 - 1 - \eta^2)\int_G |\mathbf{u}|^2. \tag{7.12}$$

If $\mathbf{u}$ were zero, then (7.3) would yield

$$\nabla(r^{\lambda-1}p) = 0.$$

Thus $r^{\lambda-1}p$ would be constant, which implies, as $\lambda \neq 1$, that p would be zero. Therefore, if we have (7.2), then $\mathbf{u} \neq 0$. So we have

$$\int_G |\nabla_s \mathbf{u}|^2 \geqq \Lambda_1 \int_G |\mathbf{u}|^2 \tag{7.13}$$

and, with (7.12) it implies that

$$3\xi^2 - 1 - \eta^2 \geqq \Lambda_1,$$

i.e.,

$$\xi \geqq \frac{1}{\sqrt{3}}(\Lambda_1 + 1 + \eta^2)^{1/2}.$$

As $\xi > -\frac{1}{2}$, it is equivalent to

$$\xi > \frac{1}{\sqrt{3}}(\Lambda_1 + 1 + \eta^2)^{1/2}. \qquad \square$$

7.3. The case when λ is real: $-\frac{1}{2} < \lambda < 1$. As $\lambda \neq 1$, we determine $\int_G p\bar{\mathbf{u}}_r$ by using (7.7), and putting that into (7.3), we get

$$\int_G |\nabla_s \mathbf{u}|^2 - \lambda(\lambda+1)|\mathbf{u}|^2 = \frac{2\lambda+1}{\lambda-1}\int_G |\nabla_s \mathbf{u}_r|^2 - (\lambda+1)(\lambda+2)|\mathbf{u}_r|^2. \tag{7.14}$$

But, with (7.13), we have

$$\begin{aligned} \int_G |\nabla_s \mathbf{u}|^2 - \lambda(\lambda+1)|\mathbf{u}|^2 &\geqq [\Lambda_1 - \lambda(\lambda+1)]\int_G |\mathbf{u}|^2 \\ &\geqq [\Lambda_1 - \lambda(\lambda+1)]\int_G |\mathbf{u}_r|^2 \end{aligned} \tag{7.15}$$

whenever

$$\Lambda_1 \geqq \lambda(\lambda+1). \tag{7.16}$$

So, with (7.15), (7.14) yields

$$\frac{2\lambda+1}{\lambda-1}\int_G |\nabla_s \mathbf{u}_r|^2 - (\lambda+1)(\lambda+2)|\mathbf{u}_r|^2 \geqq [\Lambda_1 - \lambda(\lambda+1)]\int_G |\mathbf{u}_r|^2.$$

As $\lambda-1$ is negative, this may be written in the form

$$\int_G |\nabla_s \mathbf{u}_r|^2 \leqq \phi(\lambda)\int_G |\mathbf{u}_r|^2 \tag{7.17}$$

with

$$\phi(\lambda) = (\lambda+1)(\lambda+2) - [\Lambda_1 - \lambda(\lambda+1)]\frac{1-\lambda}{2\lambda+1}. \tag{7.18}$$

Just as in [16], we introduce Definition 7.19.

DEFINITION 7.19. *Let Λ' be the minimum of $\int_G |\nabla_s v|^2$ when $v \in \mathring{H}^1(G)$, $\|v\|_{L^2(G)} = 1$ and $\int_G v = 0$.*

Formulae (7.17) and (7.8) imply that if $\phi(\lambda) < \Lambda'$, then $\mathbf{u}_r = 0$. With (7.16), this implies that $\mathbf{u} = 0$ and $p = 0$ since $\lambda \neq 1$. Therefore, we get the following lemma.

LEMMA 7.20. *Let us suppose that we have* (7.1)-(7.4) *and moreover that λ is real and smaller than one. Then, with* (7.18) *and Definition* 7.19, *we have*

$$\lambda(\lambda+1) > \Lambda_1 \quad \textit{or} \quad \phi(\lambda) \geqq \Lambda'.$$

This statement is close to what is proved in [16].

7.4. The case when $\lambda = 1$. Equations (7.6) and (7.8), respectively, give

$$\delta \mathbf{u}_r - 6\mathbf{u}_r = 0, \tag{7.21}$$

$$\int_G \mathbf{u}_r = 0. \tag{7.22}$$

Then if $\mathbf{u}_r$ is nonzero, six is an eigenvalue of δ. But since the first eigenfunction has a constant sign (cf. [4] and (19.B) in [9]), six cannot be equal to Λ_1. Therefore, if $6 < \Lambda_2$ (the second eigenvalue of δ), then $\mathbf{u}_r = 0$. So (7.5) implies that

$$\int_G |\nabla_s \mathbf{u}|^2 - 2|\mathbf{u}|^2 = 0.$$

If $2 < \Lambda_1$, then $\mathbf{u} = 0$ and p is a constant. We have just proved Lemma 7.23.

LEMMA 7.23. *Let us suppose that $\lambda = 1$ and that*

$$\Lambda_1 > 2 \quad \textit{and} \quad \Lambda_2 > 6.$$

Then the solutions of (7.2)-(7.4) *are proportional to* $(\vec{0}, 1)$.

Now, we are going to prove Lemma 7.24.

LEMMA 7.24. *If $\Lambda_1 \geqq 6$, then the pole of $\mathscr{L}_3(\lambda)^{-1}$ in $\lambda = 1$ has the order one.*

Proof. Because no confusion is possible, we drop the index 3 in $\mathscr{L}_3(\lambda)$.

Since $\mathscr{L}(1)$ is not injective, $\mathscr{L}^{-1}(\lambda)$ has a pole in $\lambda = 1$. Let us write the Laurent expansion of $\mathscr{L}^{-1}(\lambda)$ in the neighborhood of $\lambda = 1$, and the power series of $\mathscr{L}(\lambda)$:

$$\mathscr{L}(\lambda)^{-1} = \sum_{j \geqq -J} (\lambda-1)^j A_j$$

where J is the order of the pole

$$\mathscr{L}(\lambda) = \sum_{j \geqq 0} (\lambda - 1)^j \mathscr{L}^{(j)}(1)/j!$$

where $\mathscr{L}^{(j)}(\lambda)$ is the jth derivative of $\mathscr{L}$ with respect to λ. As $\mathscr{L}(\lambda)\mathscr{L}^{-1}(\lambda) = I$, we get the relation

$$\mathscr{L}(1)A_{-J} = 0, \tag{7.25}$$

and, only if $J \geqq 2$,

$$\mathscr{L}^{(1)}(1)A_{-J} + \mathscr{L}(1)A_{-J+1} = 0. \tag{7.26}$$

Thus, (7.25) implies that $A_{-J} = \Phi \cdot (0, 1)$ where Φ is a linear form, and using (7.26) we get that if

$$\mathscr{L}^{(1)}(1)(0, 1) \notin \mathscr{L}(1)D^0(G), \tag{7.27}$$

then $J = 1$. On the other hand

$$\mathscr{L}^{(1)}(1)(0, 1) = (\vec{\Psi}, 0), \tag{7.28}$$

and

$$\mathscr{L}(1)D^0(G) = (\ker \mathscr{L}(1)^*)^{\perp}. \tag{7.29}$$

But, according to [16], we have

$$\mathscr{L}(1)^*(\mathbf{u}, p) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \mathscr{L}(-2)(\mathbf{u}, -p). \tag{7.30}$$

So we search the kernel of $\mathscr{L}(-2)$, which is one-dimensional just like the kernel of $\mathscr{L}(1)$. We may suppose that a basis of ker $\mathscr{L}(-2)$ has the radial form $(\mathbf{u}, p) = (v\vec{\Psi}, p)$. According to (7.6), we have

$$p = \tfrac{1}{3}\delta v. \tag{7.31}$$

We also have

$$\begin{aligned} \operatorname{div}(r^{-2}v\vec{\Psi}) &= \langle \vec{\nabla}, r^{-3}v\vec{z} \rangle \\ &= r^{-3}v\langle \vec{\nabla}, \vec{z} \rangle + \langle \vec{z}, \vec{\nabla} \rangle (r^{-3}v) \\ &= 3r^{-3}v + (r\partial_r)(r^{-3}v) = 0. \end{aligned}$$

As a consequence, relation (7.4) is satisfied for any v (see also (7.8)). To take into account relation (7.3), we notice that

$$\begin{aligned} &\Delta(r^{-3}z_j v) = z_j \Delta(r^{-3}v) + 2\partial_j(r^{-3}v), \\ &-\Delta(r^3 v) = r^{-5}(\delta v - 6v), \\ &\partial_j(r^{-3}v) = -3z_j r^{-5}v + r^{-3}\partial_j v. \end{aligned}$$

So (7.3), which may be written as

$$-\Delta(r^{-3}z_j v) + \tfrac{1}{3}\partial_j(r^{-3}\delta v) = 0, \qquad j = 1, 2, 3,$$

is equivalent to

$$\partial_j(\delta v - 6v) = 0, \qquad j = 1, 2, 3.$$

We have just proved Lemma 7.32. □

LEMMA 7.32. *A basis of* $\ker \mathscr{L}(-2)$ *is given by* $(v\vec{\Psi}, p)$, *where* $p = \delta v/3$ *and*

(i) *If six is an eigenvalue of* δ: v *is an eigenfunction of* δ *associated with six*;

(ii) *If not,* v *is the unique solution of* $\delta v - 6v = 1$.

End of the proof of Lemma 7.4. Using (7.28)-(7.30), we may rewrite condition (7.27):

$$\int_G \langle \vec{\Psi}, v\vec{\Psi} \rangle \neq 0,$$

i.e.,

$$\int_G v \neq 0. \tag{7.33}$$

If $\Lambda_1 = 6$, v has a constant sign and (7.33) is fulfilled. If $\Lambda_1 \neq 6$, according to the assumptions of Lemma 7.24, $\Lambda_1 > 6$. Let (Λ_k, v_k) be the eigenvalues and eigenfunctions sequence of δ. Using Lemma 7.32, we have

$$v = \sum_k c_k v_k \quad \text{with } c_k = (\Lambda_k - 6)^{-1} \int_G v_k.$$

So

$$\int_G v = \sum_k (\Lambda_k - 6)^{-1} \left(\int_G v_k \right)^2,$$

which is positive since $\Lambda_1 > 6$. Thus (7.33) is true. □

8. Study of the minimum value Λ'.

8.1. Minoration of $\Lambda'(G)$. We study the minimum $\Lambda'(G)$ of $\int_G |\nabla_s v|^2$ when $v \in \mathring{H}^1(G)$, $\|v\|_{L^2(G)} = 1$ and $\int_G v = 0$ (cf. Definition 7.19) occurring in Lemma 7.20. As the extension by zero preserves the above conditions of v, we get (as in [16]):

$$\text{if } G_1 \subset G_2, \text{ then } \Lambda'(G_1) \geqq \Lambda'(G_2). \tag{8.1}$$

The minimum $\Lambda'(G)$ is reached for some function v. We recall that (Λ_k, v_k) is the eigenvalues and eigenfunctions sequence of the Laplace-Beltrami operator δ on $\mathring{H}^1(G)$. We denote

$$\gamma_k = \gamma \int_G v_k \, d\Psi \quad \text{where } \gamma = \left(\int_G d\Psi \right)^{-1/2}$$

(we suppose that $\|v_k\|_{L^2(G)} = 1$). We have

$$v = \sum_k c_k v_k \quad \text{with } \sum c_k^2 = 1, \tag{8.2}$$

$$\sum_k c_k \gamma_k = 0, \tag{8.3}$$

$$\Lambda'(G) = \sum_k \Lambda_k c_k^2. \tag{8.4}$$

If $G = S^2$, as $\gamma_1 = 1$ and the other γ_k are zero, it is obvious that (cf. [16])

$$\Lambda'(S^2) = \Lambda_2 = 2. \tag{8.5}$$

So, by (8.1) and (8.5), we get

$$\Lambda'(G) \geqq 2. \tag{8.6}$$

We will obtain further information about $\Lambda'(G)$.

LEMMA 8.7. *Let K be $(\int_G v_1\, d\Psi)^2(\int_G d\Psi)^{-1}$. Then*

$$\Lambda'(G) \geqq (1-K)\Lambda_1(G) + K\Lambda_2(G).$$

Proof. Using (8.3), we get

$$c_1^2\gamma_1^2 \leqq \left(\sum_{k\geqq 2} c_k^2\right)\left(\sum_{k\geqq 2}\gamma_k^2\right) = (1-c_1^2)(1-\gamma_1^2).$$

So, we have

$$c_1^2 \leqq 1-\gamma_1^2. \tag{8.8}$$

Equation (8.4) implies that

$$\Lambda'(G) \geqq \Lambda_1 c_1^2 + \Lambda_2(1-c_1^2).$$

Using (8.8), we get

$$\Lambda'(G) \geqq \Lambda_1(1-\gamma_1^2) + \Lambda_2\gamma_1^2.$$

And as K is exactly γ_1^2, we get the lemma. □

Now, if $\gamma_2 = 0$, instead of (8.8) we get

$$c_1^2 \leqq (1-\gamma_1^2)(1-c_2^2).$$

And, with (8.4), we have

$$\Lambda'(G) \geqq \Lambda_1 c_1^2 + \Lambda_2 c_2^2 + \Lambda_3(1-c_1^2-c_2^2).$$

Thus

$$\Lambda'(G) \geqq [(1-\gamma_1^2)\Lambda_1 + \gamma_1^2\Lambda_3](1-c_2^2) + \Lambda_2 c_2^2. \tag{8.9}$$

And, it is easy to show that, if moreover $\gamma_3, \cdots, \gamma_{N-1} = 0$, Λ_3 may be replaced by Λ_N in (8.9). So we get Lemma 8.10.

LEMMA 8.10. *If $\gamma_2, \cdots, \gamma_{N-1} = 0$, then*

$$\Lambda'(G) \geqq \min\{[(1-K)\Lambda_1 + K\Lambda_N], \Lambda_2\}.$$

8.2. The exact value of Λ' in some special cases. For $\omega \in]0, 2\pi]$ and (θ, φ) the spherical coordinates in $[0, \pi]\times[0, 2\pi[$, we denote by G_ω:

$$G_\omega = \{\Psi \in S^2 / \theta \in]0, \pi[, \varphi \in]0, \omega[\}.$$

The associated cone Γ_ω is a dihedron with interior angle ω. Since v_1 is proportional to $\sin(\pi/\omega)\varphi$, it is easy to compute the following:

$$K(G_\omega) = 8/\pi^2 \approx 0.81057. \tag{8.11}$$

The main result is Proposition 8.12.

PROPOSITION 8.12. $\Lambda'(G_\omega) = \Lambda_2(G_\omega)$.

Proof. We denote π/ω by ν. As a consequence of (18.6′) in [9] we obtain

$$\Lambda_k = \mu_k(\mu_k + 1),$$

(μ_k) being the increasing sequence of positive numbers

$$l\nu + d \quad \text{with } l \in \mathbf{N}^* \text{ and } d \in \mathbf{N}.$$

(The multiplicity of μ is given by the number of couples (l, d) providing μ.)

From (18.9) in [9], we derive that an eigenfunction associated with $\mu_k(\mu_k+1)$ with $\mu_k = l\nu + d$ has the following form:

$$v_k = \sum_{0 \leqq 2p \leqq d} \alpha_p \cos^{d-2p}\theta \cdot \sin l\nu\varphi$$

where the α_p are some constants.

As a consequence

(8.13) if $\mu_k = \nu+1$, $v_k = \alpha \cos\theta \sin\nu\varphi$ and $\gamma_k = 0$,

(8.14) if $\mu_k = 2\nu$, $v_k = \alpha \sin 2\nu\varphi$ and $\gamma_k = 0$.

(a) If $\nu \in [\frac{1}{2}, 1]$: $\mu_1 = \nu$, $\mu_2 = 2\nu$, $\mu_3 = \nu+1$, $\mu_4 = 3\nu$.
So γ_2 and γ_3 are zero; and according to (8.10) it is sufficient to prove that

$$(1-K)\nu(\nu+1)+3K\nu(3\nu+1) \geqq 2\nu(2\nu+1).$$

With (8.11), this is easy to check.

(b) If $\nu \in [1, 2]$: $\mu_1 = \nu$, $\mu_2 = \nu+1$, $\mu_3 = 2\nu$, $\mu_4 = \nu+2$.
So γ_2 and γ_3 are zero again. And we can prove that

$$(1-K)\nu(\nu+1)+K(\nu+2)(\nu+3) \geqq (\nu+1)(\nu+2). \tag{8.15}$$

Using Lemma 8.10 we get Proposition 8.12.

(c) If $\nu \geqq 2$: $\mu_1 = \nu$, $\mu_2 = \nu+1$, $\mu_3 = \nu+2$. $\gamma_2 = 0$ and as in case (b), (8.15) is true and implies Proposition 8.12 by using Lemma 8.10. □

COROLLARY 8.16. *If* $\omega \in [\pi, 2\pi]$, $\Lambda'(G_\omega) = (2\pi/\omega)(1+2\pi/\omega)$.

9. Precise regularity results in three-dimensional domains.

9.1. Strips free of poles. Let Ω be a domain in $\mathcal{O}_3(\mathbf{R}^3)$. If Ω has no vertex, it has been studied in § 6 (Theorem 6.3). If not, for each vertex x of Ω, we must check condition (3.7).

Let us assume that

$$s < \tfrac{3}{2}. \tag{9.1}$$

So, using Lemma 4.2, we have that (3.7) is equivalent to

$$\forall\lambda,\ \mathrm{Re}\,\lambda \in [-\tfrac{1}{2}, s-\tfrac{1}{2}],\ \mathscr{L}_3(\lambda) \text{ is invertible on } D^0_3(G_x). \tag{9.2}$$

We are going to determine $s(G_x)$ so that (9.2) is true for $s = s(G_x)$.

We denote

$$\xi_0 = s(G_x) - \tfrac{1}{2}.$$

As a consequence of Lemmas 7.9 and 7.20, if we have the three following conditions, for a $\xi \in]-\frac{1}{2}, 1[$:

$$\xi(\xi+1)+(2\xi+1)(\xi-1) < \Lambda_1(G), \tag{9.3}$$

$$\xi(\xi+1) < \Lambda_1(G), \tag{9.4}$$

$$\phi(\xi) < \Lambda'(G), \tag{9.5}$$

then

$$\forall\lambda,\ \mathrm{Re}\,\lambda = \xi,\ \mathscr{L}_3(\lambda) \text{ is invertible on } D^0_3(G). \tag{9.6}$$

Condition (9.4) implies (9.3), and (9.5) may be written as

$$(\xi+1)(\xi^2+6\xi+2) < (2\xi+1)\Lambda' + (1-\xi)\Lambda_1. \tag{9.7}$$

Using (8.6), $\Lambda' \geqq 2$ and $\Lambda_1 > 0$, we obtain (9.7) if

$$(\xi+1)(\xi^2+6\xi+2) \leqq 2(2\xi+1).$$

It is easy to check that for all $\xi \in [-\frac{1}{2}, 0]$.

So

$$\xi_0 \geqq 0, \quad \text{i.e.} \quad s(G) \geqq \tfrac{1}{2}. \tag{9.8}$$

For $\xi > 0$, we may use one of the following three conditions, each implying (9.7):

$$(\xi+1)(\xi^2+6\xi+2) < 2(2\xi+1)+(1-\xi)\Lambda_1, \tag{9.9}$$

$$(\xi+1)(\xi^2+6\xi+2) \leqq (2\xi+1)\Lambda', \tag{9.10}$$

$$(\xi+1)(\xi^2+6\xi+2) \leqq (\xi+1)\Lambda'+\Lambda_1 \tag{9.11}$$

(since $\Lambda' > \Lambda_1$). Each of these conditions has the form

$$\psi(\xi) \leqq 0,$$

ψ being a strictly convex function on $[-7/3, +\infty[$. To have $\psi(\xi) \leqq 0$ on $\xi \in [0, \xi_0]$, it is enough to check that

$$\psi(0) \leqq 0 \quad \text{and} \quad \psi(\xi_0) \leqq 0.$$

Using (9.9), we find Proposition 9.12, as in [16].

PROPOSITION 9.12. *$s(G)$ may be taken as $\frac{1}{2}+\mu/(\mu+4)$, where $\mu > 0$ is such that $\mu(\mu+1) = \Lambda_1(G)$.*

Now, if we consider condition (9.11) with $\Lambda' \geqq 2$, we get

$$(\xi+1)(\xi^2+6\xi) \leqq \Lambda_1,$$

and it is easy to prove the following proposition.

PROPOSITION 9.13. *If $\Lambda_1(G) \leqq 2$, $s(G)$ may be taken to be $\frac{1}{2}\,\Lambda_1(G)/8$.*

It is better than Proposition 9.12 if $\Lambda_1 \geqq 1.20$.

We notice that the right-hand side of (9.11) increases when the domain G decreases. So, we may determine ξ such that (9.11) is satisfied for G_ω (cf. 8.3), and then we are sure that (9.11) is also satisfied for all $G \subset G_\omega$. It is not difficult to check Proposition 9.14.

PROPOSITION 9.14. *If $G \subset G_\omega$ with $\omega \in]\pi, 2\pi]$, $s(G)$ may be taken as $6\pi/5\omega$.*

Here, we use Corollary 8.16:

$$\Lambda'(G_\omega) = 2\nu(\nu+1) \quad \text{and} \quad \Lambda_1(G_\omega) = \nu(\nu+1), \quad \text{with} \quad \nu = \pi/\omega.$$

In the important case when $\omega = \pi$, we have

$$\Lambda'(G_\pi) = 6$$

and we immediately see that (9.10) is true for all $\xi \leqq 1$. Therefore, we have (9.5) for $\xi < 1$ and as $\Lambda_1 = 2$, we have (9.4) also. Thus we have Proposition 9.15.

PROPOSITION 9.15. *If $G \subset G_\pi$, then we may take $s(G) = \frac{3}{2} - \varepsilon$, for all $\varepsilon > 0$.*

Finally, for $G_{\pi/2}$, $\Lambda_1 = 6$ and according to Lemma 7.24 we get Proposition 9.16.

PROPOSITION 9.16. *If $G \subset G_{\pi/2}$, $\mathscr{L}_3(\lambda)^{-1}$ has only one pole in the strip* $\operatorname{Re}\lambda \in [-\frac{1}{2}, 1]$. *That pole is $\lambda = 1$, it is simple, and* $\operatorname{Ker}\mathscr{L}(1)$ *is generated by* $(\vec{0}, 1)$. *Here let $s(G)$ be 3/2.*

9.2. Regularity results. Just as in Theorem 6.3, we make various assumptions concerning the behavior of g at the singular points of Ω, according to the value of s.

DEFINITION 9.17. Let g be in $H^{s-1}(\Omega)$ for $s \geqq 1$. g is said to be *zero at the singular points of* Ω

(i) When $s=1$: $\rho_1^{-1} g \in L^2(\Omega)$ with $\rho_1(x)$ the distance of x from $A_1(\Omega)$, the edges of Ω;

(ii) When $1<s<\frac{3}{2}$: $g=0$ on $A_1(\Omega)$;

(iii) When $s=\frac{3}{2}$: $g=0$ on $A_1(\Omega)$ and $\rho_0^{-3/2} g \in L^2(\Omega)$ where ρ_0 is the distance from $A_0(\Omega)$, the vertices of Ω;

(iv) When $s>\frac{3}{2}$: $g=0$ on $A_1(\Omega) \cup A_0(\Omega)$.

Remarks 9.18.

(1) If $g \in H^{s-1} \cap \mathring{H}^1(\Omega)$, then it is zero at the singular points of Ω.

(2) If each vertex x is in the closure of an edge, then the conditions concerning $A_0(\Omega)$ are implied by conditions concerning $A_1(\Omega)$; such is the case when Ω is a polyhedron.

This may be proved as is (AC.3) in [9].

DEFINITION 9.19. For each vertex $x \in A_0(\Omega)$, we denote by s_x the best value of $s(G_x)$ drawn from Propositions 9.12-9.16.

THEOREM 9.20. *Let* $\Omega \in \mathcal{O}_3(\mathbf{R}^3)$. *Let* $(\vec{u}, p) \in D_3^0(\Omega)$ *be the solution of* (1.1) *with* $(\vec{f}, g) \in E_3^s(\Omega)$.

(a) *If* $s<1$, $s \leqq \min_{x \in A_0(\Omega)} s_x$ *and* $s < \inf_{x \in A_1(\Omega)} \xi_1(\omega_x)$, *then* $(\vec{u}, p) \in D_3^s(\Omega)$.

(b) *If* $s \geqq 1$, $s \leqq \min_{x \in A_0(\Omega)} s_x$, $s < \inf_{x \in A_1(\Omega)} \pi/\omega_x$ *and moreover* g *is zero at the singular points of* Ω, *then* $(\vec{u}, p) \in D_3^s(\Omega)$.

For $s<1$, it is a straightforward consequence of Theorem 3.6, as in the case of Theorem 6.3.

For $s \geqq 1$, in view of Theorems 3.6 and 6.3, and the methods of proofs of [9], it is enough to note that the Mellin transform $(\vec{F}(\lambda), G(\lambda))$ of $(r^2 \vec{f}, rg)$, after localization in the neighborhood of any vertex, is defined up to $\operatorname{Re} \lambda \leqq s - \frac{1}{2}$, with values in $F_3^s(G_x)$.

Now, we derive from Theorem 9.20 the statements (1.6)-(1.12).

Since $s_x > \frac{1}{2}$ and $\xi_1(\omega_x) \geqq \frac{1}{2}$, (1.6) is a straightforward consequence of Theorem 9.20a.

In situation (1.7), the openings of the edges are $\pi/2$ or $3\pi/2$. $\xi_1(3\pi/2) > 0.544$, so $0.544 < \inf_{x \in A_1(\Omega)} \xi_1(\omega_x)$. If x is a vertex of Ω, then the associated cone Γ_x is an octant or the complementary of an octant. In the first case, Proposition 9.16 yields $s(G_x) = \frac{3}{2}$. In the last case, we note that Γ_x is included in the revolution cone

$$\Gamma = \{z \in \mathbf{R}^3 / \theta \in [0, \theta_0[\}$$

with $\theta_0 = \pi - \theta_1$, $\theta_1 = \arcsin(1/\sqrt{3})$; so $\theta_0 \approx 144.74°$. As a consequence of 18.D in [9], we find that $\Lambda_1(G_x) = \mu(\mu+1)$ with $\mu > 0.35$. By Proposition 9.12 we get

$$s(G_x) > \frac{1}{2} + \frac{0.35}{4.35} \approx 0.580 > 0.544.$$

When Ω is convex, we have $\pi/\omega_x > 1$ for any $x \in A_1(\Omega)$, and by Proposition 9.15 $s_x > 1$. Thus, we get (1.8).

If, for each $x \in A_1(\Omega)$, $\pi/\omega_x \geqq \frac{3}{2}$, as by Proposition 9.15 we have $s_x \geqq \frac{3}{2} - \varepsilon$, then for all $\varepsilon > 0$, we get (1.9).

In the situation (1.10), $\min_{x \in A_1(\Omega)} \pi/\omega_x > \frac{3}{2}$ and, for each vertex x, owing to Proposition 9.16, we have $s_x = \frac{3}{2}$.

In the situations (1.11) or (1.12), we have: $\min_{x \in A_1(\Omega)} \pi/\omega_x = 2$ and there is no vertex.

Appendix. Behavior of $\mathcal{L}_2(\lambda)^{-1}$ in the neighborhood of positive integer numbers, on the domain $]0, 2\pi[$ (the model for a crack). Given suitable changes of functions (see [11] and [5]), the problem

$$\mathcal{L}_2(\lambda)(u_1, u_2, p) = (f_1, f_2, g) \text{ with } u_1, u_2 \in \mathring{H}^1(]0, 2\pi[)$$

is equivalent to the other:

$$\text{(A1)} \qquad \begin{cases} u'' + (\lambda+1)^2 u + (1-\lambda) q = l_1, \\ (\lambda-1)u' + (1-\lambda^2)v + q' = l_2, \\ (1+\lambda)u + v' = l_3, \end{cases}$$

$$\text{(A2)} \qquad u, v \in \mathring{H}^1(]0, 2\pi[).$$

As in the above references, it can be proved that (A1) is holomorphically solvable. To solve (A1)-(A2), it remains to solve (A1) with $l_j = 0$, and for any $(c_1, c_2, c_3, c_4) \in \mathbf{C}^4$

$$\text{(A3)} \qquad \begin{aligned} v(0) &= \gamma_1, \qquad v(2\pi) = \gamma_2, \\ u(0) &= \gamma_3, \qquad u(2\pi) = \gamma_4. \end{aligned}$$

Problem (A1) with the zero right-hand side is equivalent to (A4)-(A6):

$$\text{(A4)} \qquad u = -v'(1+\lambda)^{-1},$$

$$\text{(A5)} \qquad q = (v^{(3)} + (1+\lambda)^2 v')(1-\lambda^2)^{-1},$$

$$\text{(A6)} \qquad v^{(4)} + 2(1+\lambda^2)v^{(2)} + (1-\lambda^2)^2 v = 0.$$

For $\operatorname{Re}\lambda > 0$, a basis of solutions of (A6) is given by

$$v_1 = \sin(\lambda+1)\theta, \quad v_2 = \frac{\sin(\lambda-1)\theta}{\lambda-1}, \quad v_3 = \cos(\lambda+1)\theta, \quad v_4 = \cos(\lambda-1)\theta.$$

Let $M(\lambda)$ be the four $\times$ four matrix, the columns of which are

$${}^t[v_j(0), v_j(2\pi), -v_j'(0)(1+\lambda)^{-1}, -v_j'(2\pi)(1+\lambda)^{-1}].$$

The solvability of (A3), (A4), and (A6) is equivalent to finding $\alpha = {}^t(\alpha_1, \cdots, \alpha_4)$ such that

$$\text{(A7)} \qquad M(\lambda)\alpha = \gamma$$

where $\gamma = {}^t(\gamma_1, \cdots, \gamma_4)$. Then the solution of (A3), (A4), and (A6) is

$$\text{(A8)} \qquad v = \sum \alpha_j v_j \quad \text{and} \quad u = -v'(\lambda+1)^{-1}.$$

The determinant of $M(\lambda)$ is

$$4\sin^2 2\pi\lambda(\lambda-1)^{-1}(\lambda+1)^{-2}.$$

So, in $\lambda = 1$, $M(\lambda)^{-1}$ has a simple pole. On the other hand, when λ is an integer number and $\lambda \geqq 2$, it is easy to see that the first and the third rows (respectively, the second and the fourth) are equal; then the cofactors of $M(\lambda)$ are zero and the pole is simple again.

It is obvious that, for integer λ

$$\text{(A9)} \qquad \dim \operatorname{Ker} M(1) = 1 \quad \text{and} \quad \dim \operatorname{Ker} M(\lambda) = 2 \quad \text{when } \lambda \geqq 2.$$

To deduce the properties of $\mathscr{L}_2(\lambda)^{-1}$, we must take (A5) into account. For $\lambda \geqq 2$, it is holomorphic. It remains to study the case when $\lambda = 1$.

For $v = v_1$, or $v = v_3$, (A5) yields $q = 0$.

For $v = v_4$, (A5) yields $q = ((\lambda+1)^2 - (\lambda-1)^2) \sin(\lambda-1)\theta/(\lambda+1)$.

For $\lambda = 1$, it is zero again.

For $v = v_2$, (A5) yields $q_2 = ((\lambda+1)^2 - (\lambda-1)^2) \cos(\lambda-1)\theta/(1-\lambda^2)$.

With (A8), we have $q = \alpha_2 q_2$.

As q_2 has a simple pole in $\lambda = 1$, it remains to state that α_2 is holomorphic. That arises from the structure of $M(\lambda)$: the matrix $M_2(\lambda)$ obtained by removing the second column of $M(\lambda)$ has its rank equal to two, and the corresponding cofactors balance the determinant of $M(\lambda)$.

So, for any $\lambda \in \mathbf{N}$, $\lambda \geqq 1$, $\mathscr{L}(\lambda)^{-1}$ has a simple pole, and with (A9), it is clear that

$$\dim \operatorname{Ker} \mathscr{L}_2(\lambda) = d(\lambda).$$

With Lemma 4.5, we get that $\mathscr{S}_2$ is injective modulo polynomial on $S^\lambda(\mathbf{R}^2 \backslash \mathbf{R}^+)$.

REFERENCES

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

[2] C. BERNARDI AND G. RAUGEL, *Méthodes mixtes pour les équations de Navier-Stokes dans un ouvert polygonal plan*, Rapport interne du laboratoire d'analyse numérique, Université de Paris VI, 1980.

[3] ———, *Méthodes d'éléments finis mixtes pour les équations de Stokes et de Navier-Stokes dans un polygone non convexe*, Calcolo, 18 (1981), pp. 255-291.

[4] R. COURANT AND D. HILBERT, *Methods of mathematical physics*, Vol. 1, Interscience, New York, 1953.

[5] M. DAUGE, *Opérateur de Stokes dans des espaces de Sobolev à poids sur des domaines anguleux*, Canad. J. Math., 34 (1982), pp. 853-882.

[6] ———, *Régularités et singularités des solutions de problèmes aux limites elliptiques sur des domaines singuliers de type à coins*, Thèse d'état. Nantes, France, 1986.

[7] ———, *Etude de l'opérateur de Stokes dans un polygone: Régularité, singularité et théorèmes d'indice*, Thèse de 3ième cycle, Nantes, France, 1980.

[8] ———, *Problèmes aux limites dans des domaines à coins: Cas limite et démonstration d'un résultat de régularité*, C. R. Acad. Sci. Paris, Sér. I, 304 (1987), pp. 579-582.

[9] ———, *Elliptic boundary value problems on corner domains*, Lecture Notes in Mathematics 1341, Springer-Verlag, Berlin, New York, 1988.

[10] P. GRISVARD, *Singularités des solutions du problème de Stokes dans un polygone*, Université de Nice, Nice, France, 1979.

[11] R. B. KELLOGG AND J. E. OSBORN, *A regularity for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397-431.

[12] V. A. KONDRAT'EV, *Boundary value problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227-313. (In English) Trudy Moskov. Mat. Obshch., 16 (1967), pp. 209-292. (In Russian.)

[13] R. LOZI, *Résultats numériques de régularité du problème de Stokes et du laplacien itéré dans un polygone*, RAIRO Anal. Numér., 12 (1978), pp. 267-282.

[14] L. V. MASLOVSKAYA, *Behavior of solutions of boundary value problems for the biharmonic equation in domains with angular points*, Differentsial'nya Uraveniya, 19 (1983), pp. 2172-2175. (In Russian.)

[15] V. G. MAZ'JA AND B. A. PLAMENEVSKII, *First boundary value problem for the equation of hydrodynamics in a domain with a piecewise-smooth boundary*, J. Soviet. Math., (1983), pp. 777-782. (In English) Zap. Nauchn. Sem. Leningrad, Otdel. Math. Inst. Steklov. (LOMI), 96 (1980), pp. 179-186. (In Russian.)

[16] ———, *On properties of solutions of three-dimensional problems of elasticity theory and hydrodynamics in domains with isolated singular points*, Amer. Math. Soc. Trans. (2), 123 (1984), pp. 109-123. (In English.) Translated from: Dinamika Sploshnoi Sredy Vyp., 50 (1981), pp. 99-120. (In Russian.)

[17a] ———, *First boundary value problem for the classical equations of physics in domains with piecewise-smooth boundary* (I), Z. Anal. Anwendungen, 2 (1983), pp. 335-359. (In Russian.)

[17b] V. G. MAZ'JA AND B. A. PLAMENEVSKII, *First boundary value problem for the classical equations of physics in domains with piecewise-smooth boundary* (II), Z. Anal. Anwendungen, 2 (1983), pp. 523–551. (In Russian.)

[18] M. MERIGOT, *Régularité de la solution du problème de Stokes sur un polygone*, Preprint, Nice, France, 1975.

[19] J. E. OSBORN, *Regularity of solutions of the Stokes problem in a polygonal domain*, in Numerical Solutions of Partial Differential Equations III, Synspade 1975, Academic Press, New York, 1976, pp. 393–411.

[20] M. POGU AND G. TOURNEMINE, *Equation de Navier–Stokes et condition de Kutta–Joukowski généralisée*, preprint INSA Rennes, 1985.

[21] J. B. SEIF, *On the Green's function for the biharmonic equation on an infinite wedge*, Trans. Amer. Math. Soc., 182 (1973), pp. 241–260.

[22] R. TEMAM, *Navier–Stokes Equations, Theory and Numerical Analysis*, North-Holland, Amsterdam, 1977.

[23] H. TRIEBEL, *Interpolation Theory. Function Spaces. Differential Operators*, North-Holland, Amsterdam, 1978.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

007

CAUCHY FORMULAE FOR FUNCTIONS ANALYTIC OF ORDER TWO ON C^1 DOMAINS WITH APPLICATIONS TO ELASTOSTATICS AND HYDROSTATICS*

JONATHAN COHEN†

Abstract. This paper gives Cauchy-type formulae for functions analytic of order two on C^1 domains obtained from the solutions of corresponding biharmonic problems. Functions analytic of order two are shown to be potentials for the solutions of systems of linear partial differential equations in two-dimensional elastostatics and hydrostatics. When combined with Cauchy formulae, integral representations are obtained for the traction problem and the linearized Stokes problem that are valid even for C^1 domains.

Key words. Cauchy formulae, analytic of order two, traction problem, stationary Stokes problem, biharmonic equation, layer potentials

AMS(MOS) subject classifications. 30, 35, 42, 45, 73, 76

1. Introduction. In this paper, we give Cauchy-type formulae for functions analytic of order two on C^1 domains obtained from the solutions of corresponding biharmonic problems. We show how functions analytic of order two are potentials for the solutions of systems of linear partial differential equations in two-dimensional elastostatics and hydrostatics. When combined with Cauchy formulae we obtain integral representations for the traction problem and the linearized Stokes problem that are valid even for C^1 domains.

The traction problem in elastostatics is to determine the components of stress and displacement from a system of partial differential equations satisfied within a domain and the forces applied on the boundary. Airy showed [2] that in the absence of body forces this problem could be reduced to finding a scalar biharmonic potential called the stress function. The stationary Stokes problem in hydrostatics is the following. Solve a system of partial differential equations satisfied by the velocity and pressure that can likewise be reduced to a biharmonic problem. (See Mikhlin [11, pp. 176–178] for an outline of these reductions.)

In [7] and [8] Cohen and Gosselin obtained solutions to the following biharmonic problems on C^1 domains in $\mathbb{R}^2$:

$$\begin{aligned} &\Delta^2 u = 0 \quad \text{in } \Omega, \\ &\nabla u|_{\partial\Omega} = \mathbf{g} \quad \text{where } \int \mathbf{g}\cdot\mathbf{T}\,ds = 0, \end{aligned} \tag{1.1}$$

$\mathbf{T}$ is the unit tangent vector and $\mathbf{g} \in L^p \times L^p(\partial\Omega)$, $1 < p < \infty$

$$\Delta^2 u = 0, \qquad (u_{xx}x_s + u_{xy}y_s,\ u_{xy}x_s + u_{yy}y_s) = \boldsymbol{\varphi} \tag{1.2}$$

where $\boldsymbol{\varphi} = (\varphi, \psi) \in L^q \times L^q(\partial\Omega)$ and $\int_{\partial\Omega} \varphi = \int_{\partial\Omega} \psi = \int_{\partial\Omega} x\varphi + y\psi\, ds = 0$.

The first of these problems involves Dirichlet-type boundary conditions and solves the stationary Stokes problem. The second involves adjoint or Neumann-type boundary conditions and solves the traction problem.

The solutions are given by potentials that are modified versions of the multiple layer potentials introduced by Agmon in [1]. The analysis at the boundary is obtained

* Received by the editors July 21, 1986; accepted for publication (in revised form) March 21, 1988.
† Department of Mathematics, De Paul University, Chicago, Illinois 60614.

for C^1 domains via an application of Calderon's theorem on the Cauchy integral along Lipschitz curves [5].

In practice it can be difficult to obtain the appropriate biharmonic potentials. Muskhelishvili's book on elasticity [12, Chap. 5] shows how assuming the existence of Airy's stress function leads to the reformulation of the traction problem as a boundary value problem for a system of analytic functions. The crucial point in the introduction of analytic functions is that biharmonic functions can be represented as $\operatorname{Re}\{\bar{z}f(z)+g(z)\}$ where f and g are analytic. If we define $\bar{\partial}=\partial_x+i\partial_y$, a simple calculation shows that $\bar{\partial}^2(\bar{z}f(z)+g(z))=0$. This suggests a connection between elasticity and the $\bar{\partial}^2$ equation.

In fact, this connection is neither new nor surprising. In the 1920s Burgatti, in [3] and [4], studied solutions of $\bar{\partial}^n f=0$ and introduced solutions of the equation $\bar{\partial}^2 f=0$ into the equations of elasticity to obtain Kolosoff's formula for the complex displacement [4, pp. 90–91].

In this paper we look at the complex function $U+i\tilde{U}$ where U is the Airy stress function and $\tilde{U}$ is a biharmonic conjugate of U in the sense that $\bar{\partial}^2(U+i\tilde{U})=0$. We show how the displacements can be computed, up to a rigid infinitesimal deformation, as a linear combination of the first derivatives of the stress function and its biharmonic conjugate $\tilde{U}$. We then show how the layer potential representation of the solution to the biharmonic reformulation of the traction problem can be extended to a "Cauchy type" formula that automatically produces the biharmonic conjugate of the stress function.

This procedure is simpler than the method outlined on pages 106–109 of Muskhelishvili [12]. Furthermore, for half planes with any orientation, the layer potential solutions reduce to Poisson integrals of the boundary forces.

2. The $\bar{\partial}^2$ equation. Functions satisfying the equation $\bar{\partial}^2\psi=0$ are called analytic of order two where $\bar{\partial}=\partial_x+i\partial_y$ and ψ is complex valued. In this section we review some of the basic properties of these functions, some of which are discussed in a more general context in the articles by Burgatti [3] and [4].

We let ∂ denote the operator $\partial_x-i\partial_y$ and assume that ψ is complex valued and satisfies $\bar{\partial}^2\psi(z)=0$ in a domain Ω. If we write $\psi=U+i\tilde{U}$ where U and $\tilde{U}$ are smooth real-valued functions and observe that $\bar{\partial}^2=\partial_{xx}-\partial_{yy}+2i\partial_{xy}$, then $\bar{\partial}^2\psi=0$ implies

$$U_{xx}-U_{yy}-2\tilde{U}_{xy}+i(2U_{xy}+\tilde{U}_{xx}-\tilde{U}_{yy})=0. \tag{2.1}$$

Equating real and imaginary parts, we obtain the following system of second-order partial differential equations:

$$U_{xx}-U_{yy}=2\tilde{U}_{xy}, \tag{2.2}$$

$$2U_{xy}=-(\tilde{U}_{xx}-\tilde{U}_{yy}), \tag{2.3}$$

which is analogous to the Cauchy-Riemann equations.

Since the Laplace operator can be factored as $\Delta=\bar{\partial}\partial$ we observe the following:

$$\Delta U+i\Delta\tilde{U} \quad \text{is analytic} \tag{2.4}$$

since $\bar{\partial}(\Delta U+i\Delta\tilde{U})=\partial\bar{\partial}^2(U+i\tilde{U})=0$;

$$\Delta^2 U=\operatorname{Re}\Delta^2(U+i\tilde{U})=\operatorname{Re}\partial^2\bar{\partial}^2(U+i\tilde{U})=0; \tag{2.5}$$

$$\Delta^2\tilde{U}=\operatorname{Im}\Delta^2(U+i\tilde{U})=\operatorname{Im}\partial^2\bar{\partial}^2(U+i\tilde{U})=0. \tag{2.6}$$

Thus U and $\tilde{U}$ are biharmonic and since $\Delta U + i\Delta\tilde{U}$ is analytic, the pair of functions ΔU and $\Delta\tilde{U}$ satisfy the Cauchy-Riemann equations:

$$(\Delta U)_x = (\Delta\tilde{U})_y, \tag{2.7}$$

$$(\Delta U)_y = -(\Delta\tilde{U})_x. \tag{2.8}$$

The system of real second-order equations (2.2) and (2.3) together with (2.8) and (2.9) will be referred to as the biharmonic Cauchy-Riemann equations.

Analytic functions of order two can be represented as

$$\psi(z) = \frac{\bar{z}}{2} f(z) + g(z) \tag{2.9}$$

where f and g are analytic. It follows that $f(z) = \bar{\partial}\psi(z)$ and $g(z) = \psi(z) - (\bar{z}/2)\bar{\partial}\psi(z)$. We then write

$$\psi(z) = \frac{\bar{z}}{2}\bar{\partial}\psi(z) + \left(\psi(z) - \frac{\bar{z}}{2}\bar{\partial}\psi(z)\right), \tag{2.10}$$

and applying Cauchy's formula to the analytic functions $\bar{\partial}\psi(z)$ and $\psi(z) - (\bar{z}/2)\bar{\partial}\psi(z)$, we obtain the Cauchy-type representation

$$\psi(z) = \frac{1}{2\pi i}\int_\gamma \frac{\psi(w)}{w - z}\, dw - \frac{1}{4\pi i}\int_\gamma \bar{\partial}\psi(w)\frac{\bar{w} - \bar{z}}{w - z}\, dw \tag{2.11}$$

where γ is a simple closed contour in Ω containing z.

Formula (2.11) appears (except for a small error) on page 88 of Burgatti [4], and we will refer to (2.11) as Burgatti's formula.

3. The traction problem. The traction problem is to obtain the elastostatic state of a thin plate from the forces at the edge. In this section we assume that Ω is a bounded C^1 simply connected domain in $\mathbb{R}^2$. We let $S = \left[\begin{smallmatrix} A & B \\ B & C \end{smallmatrix}\right]$ denote the stress tensor and let u, v denote the components of displacement. In the absence of body forces the equilibrium equations are

$$A_x + B_y = 0, \qquad B_x + C_y = 0 \tag{3.1}$$

and the equations relating displacement and stress are

$$A = (\lambda + 2\mu)u_x + \lambda v_y, \qquad B = \mu(u_y + v_x), \qquad C = \lambda u_x + (\lambda + 2\mu)v_y. \tag{3.2}$$

If we let (X_n, Y_n) denote the normal stress along the boundary $\partial\Omega$ and (x_s, y_s) the unit tangent vector on $\partial\Omega$, then the traction problem is to find components of stress A, B, C and displacements u, v satisfying (3.1) and (3.2) in Ω and for which

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix}\begin{bmatrix} y_s \\ -x_s \end{bmatrix} = \begin{bmatrix} X_n \\ Y_n \end{bmatrix} \quad \text{on } \partial\Omega.$$

Airy showed that the equilibrium equations (3.1) imply the existence of a function w such that $A = w_{yy}$, $B = -w_{xy}$, and $C = w_{xx}$. Furthermore, substituting the second partials of w for A, B, and C into (3.2), it easily follows that $\Delta^2 w = 0$. This means that the traction problem can be reformulated as the biharmonic problem:

$$\begin{aligned} &\Delta^2 w = 0, \\ &\begin{bmatrix} w_{yy} & -w_{xy} \\ -w_{xy} & w_{xy} \end{bmatrix}\begin{bmatrix} y_s \\ -x_s \end{bmatrix} = \begin{bmatrix} X_n \\ Y_n \end{bmatrix} \quad \text{in } \partial\Omega. \end{aligned} \tag{3.3}$$

A quick inspection shows that this is problem (1.2) in the Introduction. The solution w is called the Airy stress function.

To complete the solution to the traction problem it is necessary to compute the displacements from the stress function. This procedure, as outlined in Muskhelishvili [12, pp. 106–109], is somewhat complicated. We will now show how, if a biharmonic conjugate can be found for the stress function, the computation can be simplified considerably.

If we substitute the appropriate second partials of w for A and C in the first and third equations of (3.1) and solve for u_x and v_y we get

$$u_x = \frac{-\lambda w_{xx} + (\lambda + 2\mu) w_{yy}}{4\mu(\lambda + \mu)}, \tag{3.4}$$

$$v_y = \frac{(\lambda + 2\mu) w_{xx} - \lambda w_{yy}}{4\mu(\lambda + \mu)}. \tag{3.5}$$

We now assume there exists a function $\tilde{w}$ such that $\bar{\partial}^2(w + i\tilde{w}) = 0$. Using the biharmonic Cauchy-Riemann equations we can substitute $w_{xx} - 2\tilde{w}_{xy}$ for w_{yy} in (3.4) and $w_{yy} + 2\tilde{w}_{xy}$ for w_{xx} in (3.5) to get

$$u_x = \frac{\partial}{\partial x}\left(\frac{\mu w_x - (\lambda + 2\mu)\tilde{w}_y}{2\mu(\lambda + \mu)}\right), \tag{3.6}$$

$$v_y = \frac{\partial}{\partial x}\left(\frac{\mu w_y + (\lambda + 2\mu)\tilde{w}_x}{2\mu(\lambda + \mu)}\right). \tag{3.7}$$

This is the main point. The introduction of the biharmonic conjugate $\tilde{w}$ enables us to integrate u_x and v_y to obtain

$$u = \frac{\mu w_x - (\lambda + 2\mu)\tilde{w}_y}{2\mu(\lambda + \mu)} + F_1(y), \tag{3.8}$$

$$v = \frac{\mu w_y + (\lambda + 2\mu)\tilde{w}_x}{2\mu(\lambda + \mu)} + F_2(x). \tag{3.9}$$

Substituting for u and v in (3.2), we have

$$\begin{aligned} \frac{-1}{\mu} w_{xy} &= (u_y + v_x) \\ &= \frac{\mu w_{xy} - (\lambda + 2\mu)\tilde{w}_{yy}}{2\mu(\lambda + \mu)} + F_1'(y) + \frac{\mu w_{xy} - (\lambda + 2\mu)\tilde{w}_{xx}}{2\mu(\lambda + \mu)} + F_2'(x) \\ &= \frac{2\mu w_{xy} - (\lambda + 2\mu)(\tilde{w}_{yy} - \tilde{w}_{xx})}{2\mu(\lambda + \mu)} + F_1'(y) + F_2'(x), \end{aligned} \tag{3.10}$$

which, by the biharmonic Cauchy-Riemann equations,

$$\begin{aligned} &= \frac{2\mu w_{xy} - (\lambda + 2\mu)(2w_{xy})}{2\mu(\lambda + \mu)} + F_1'(y) + F_2'(x) \\ &= -\frac{1}{\mu} w_{xy} + F_1'(y) + F_2'(x). \end{aligned}$$

Thus we have $0 = F_1'(y) + F_2'(x)$. This implies that $F_1'(y) = -F_2'(x) = \varepsilon$ and so $F_1(y) = \varepsilon y + \tau$, $F_2(x) = -\varepsilon x + \sigma$.

The choice of biharmonic conjugate $\tilde{w}$ was arbitrary. However, it follows from Lemma 13.1 of Agmon [1] that if $\bar{\partial}^2(w+i\tilde{w})=0$ and $\bar{\partial}^2(w+i\tilde{w}_1)=0$, then $\tilde{w}_1-\tilde{w}=\alpha x+\beta y+\frac{1}{2}(x^2+y^2)+\delta$. Substituting $\tilde{w}_1$ for $\tilde{w}$ in (3.8) and (3.9), we get a displacement (u_1, v_1) that differs from (u, v) by

$$u_1-u=\frac{-(\lambda+2\mu)}{2\mu(\lambda+\mu)}(\gamma y+\beta), \tag{3.11}$$

$$v_1-v=\frac{(\lambda+2\mu)}{2\mu(\lambda+\mu)}(\gamma x+\alpha). \tag{3.12}$$

This means that two distinct "biharmonic conjugates" give rise to displacements that differ by, at worst, an infinitesimal rigid displacement. In other words, we have introduced no new pure deformation by computing the displacement from the derivatives of $\tilde{w}_1$ rather than $\tilde{w}$. Up to an infinitesimal rigid displacement we have the formula

$$(u, v)=\left(\frac{\mu w_x-(\lambda+2\mu)\tilde{w}_y}{2\mu(\lambda+\mu)}, \frac{\mu w_y+(\lambda+2\mu)\tilde{w}_x}{2\mu(\lambda+\mu)}\right). \tag{3.13}$$

4. The stationary Stokes problem. The stationary Stokes problem in hydrostatics has the formulation in a domain Ω:

$$\begin{aligned} &\Delta\mathbf{u}=\nabla p \quad \text{in } \Omega \\ &\operatorname{div}\mathbf{u}=0 \quad \text{in } \Omega, \\ &\mathbf{u}|_{\partial\Omega}=\mathbf{f} \end{aligned} \tag{4.1}$$

where $\mathbf{u}=(u, v)$ is the velocity of the fluid, p is the pressure, and $\mathbf{f}=(f_1, f_2)$ is the velocity at the boundary. The second equation, $\operatorname{div}\mathbf{u}=0$, implies there exists a function Φ satisfying $\nabla\Phi=(-v, u)$. It then follows from substitution for u and v in the first equation that $\Delta^2\Phi=\operatorname{div}(\Delta\Phi_x, \Delta\Phi_y)=-p_{xy}+p_{xy}=0$. The Stokes problem then has the biharmonic formulation:

$$\begin{aligned} &\Delta^2\Phi=0 \quad \text{in } \Omega, \\ &\nabla\Phi|_{\partial\Omega}=(-f_2, f_1). \end{aligned} \tag{4.2}$$

It remains to obtain the pressure p from the solution Φ of (4.2). If we assume there exists a $\tilde{\Phi}$ such that $\bar{\partial}^2(\Phi+i\tilde{\Phi})=0$, then by the second part of the biharmonic Cauchy-Riemann equations,

$$(\Delta\tilde{\Phi})_x=-(\Delta\Phi)_y=-\Delta u=-p_x, \qquad (\Delta\tilde{\Phi})_y=(\Delta\Phi)_x=-\Delta v=-p_y. \tag{4.3}$$

Hence $\nabla(-\Delta\tilde{\Phi})=\nabla p$ so that $-\Delta\tilde{\Phi}$ differs from the pressure by a constant. If a second biharmonic conjugate $\tilde{\Phi}_1$ is used, then $\Delta\tilde{\Phi}_1-\Delta\tilde{\Phi}$ is a constant.

It is clear that this same calculation shows that any harmonic conjugate of $\Delta\Phi$ will suffice to give the pressure. However, we will point out that the layer potential solution of (4.2) automatically produces a biharmonic conjugate $\tilde{\Phi}$ so that no additional integrations are necessary to find the pressure.

5. The biharmonic results. We assume that Ω is a bounded, simply connected C^1 domain in R^2 with boundary $\partial\Omega$. We next introduce the following spaces of boundary data.

DEFINITION 5.1.

$$C_p=\left\{\mathbf{g}=(g, h)\in(L^p\times L^p)(\partial\Omega): \int_{\partial\Omega} g\,dx+h\,dy=0\right\}.$$

DEFINITION 5.2. $(L^p \times L^p)_0(\partial\Omega) = \{\boldsymbol{\varphi} = (\varphi, \psi) \in (L^q \times L^q)(\partial\Omega)$: $\int_{\partial\Omega} \nabla w \cdot \boldsymbol{\varphi}\, ds = 0$ for all $w(x, y) = \alpha x + \beta y + \gamma(x^2 + y^2) + \delta$ and $1 < q < \infty\}$.

Let $\tilde{F}(x, y) = (-1/4\pi)\{(x^2 + y^2) \arg(x + iy) - xy\}$ for some particular choice of the argument. In what follows we will let X denote points in the domain Ω and P, and Q will denote points on the boundary.

DEFINITION 5.3. For $Q \in \partial\Omega$ and $X \in \Omega$ we define the boundary differential operator $\mathbf{L} = \mathbf{L}_Q$ by

$$\mathbf{L}v(X) = (L_1 v(X), L_2 v(X))$$

where

$$L_1 v(X) = v_{xx}(X)x_s(Q) + v_{xy}(X)y_s(Q), \qquad L_2 v(X) = v_{xy}(X)x_s(Q) + v_{yy}(X)y_s(Q) \tag{5.4}$$

with $(x(s), y(s))$ being the arclength parameterization of $\partial\Omega$ and $(x_s(Q), y_s(Q))$ being the unit tangent at Q.

DEFINITION 5.5. For $\mathbf{g} \in C_p$ we define the modified multiple layer potential by

$$u_m(\mathbf{g}: X) = \int_{\partial\Omega} g(Q)L_1\tilde{F}(X - Q) + h(Q)L_2\tilde{F}(X - Q)\, ds(Q). \tag{5.6}$$

For $\boldsymbol{\varphi} \in (L^q \times L^q)_0(\partial\Omega)$ we define the modified lower-order potential by

$$v_m(\boldsymbol{\varphi}; X) = \int_{\partial\Omega} \varphi(P)\tilde{F}_x(P - X) + \psi(P)\tilde{F}_y(P - X)\, ds(P). \tag{5.7}$$

For $X \notin \partial\Omega$ we can differentiate (5.6) and (5.7) under the integral signs to get

$$\nabla u_m = \int_{\partial\Omega} \mathbf{g}(Q)l(X, Q)\, ds(Q) \tag{5.8}$$

where

$$l(X, Q) = \begin{bmatrix} \partial_x^X L_1\tilde{F}(X - Q) & \partial_y^X L_1\tilde{F}(X - Q) \\ \partial_x^X L_2\tilde{F}(X - Q) & \partial_y^X L_2\tilde{F}(X - q) \end{bmatrix} \tag{5.9}$$

and

$$\mathbf{L}v_m(X)^T = \int_{\partial\Omega} l(X, P, Q)\boldsymbol{\varphi}(P)^T\, ds(P) \tag{5.10}$$

where

$$l(X, P, Q) = \begin{bmatrix} \partial_x^P L_1\tilde{F}(P - X) & \partial_y^P L_1\tilde{F}(P - X) \\ \partial_x^P L_2\tilde{F}(P - X) & \partial_y^P L_2\tilde{F}(P - X) \end{bmatrix}, \tag{5.11}$$

and the superscript T denotes the transpose of a row vector. Note that the dependence of l on Q is built into the definition of L_1 and L_2.

DEFINITION 5.12. For $P \neq Q$, $P, Q \in \partial\Omega$, we can define the matrix kernels in (5.9) and (5.11) by letting $X = Q$. Both kernels are then the same and we call them $l(P, Q)$. We define the operators

$$\mathscr{L}_\varepsilon \mathbf{g}(P) = \int_{|P-Q|>\varepsilon} \mathbf{g}(Q, Q)l(P, Q)\, ds(Q) \tag{5.13}$$

and

$$\mathscr{L}_\varepsilon^* \boldsymbol{\varphi}(P) = \int_{|P-Q|>\varepsilon} l(P, Q)\boldsymbol{\varphi}(P)^T\, ds(P). \tag{5.14}$$

We tentatively define the operators $\mathscr{L}\mathbf{g}(P) = \lim_{\varepsilon\to 0} \mathscr{L}_\varepsilon \mathbf{g}(P)$ and $\mathscr{L}^*\boldsymbol{\varphi}(Q) = \lim_{\varepsilon\to 0} \mathscr{L}_\varepsilon^*\boldsymbol{\varphi}(Q)$.

THEOREM 5.15. *For $\mathbf{g}\in C_p$, we have the following*:

(i) *$\mathscr{L}\mathbf{g}(P)$ exists almost everywhere with respect to arclength, $\mathscr{L}$ is bounded from C_p to itself in the $(L^p\times L^p)(\partial\Omega)$ norm and in fact is compact from C_p to itself.*

(ii) *The nontangential*

$$\lim_{X\to P} \nabla u_m(X) = \begin{cases} (I+\mathscr{L})\mathbf{g}(P), & X\in\Omega, \\ (-I+\mathscr{L})\mathbf{g}(P), & X\notin\bar{\Omega} \end{cases}$$

for almost every $P\in\partial\Omega$,

(iii) *$(I+\mathscr{L})^{-1}$ exists on C_p and $(-I+\mathscr{L})^{-1}$ exists on the space $(-I+\mathscr{L})(C_p)$.*

COROLLARY 5.16. *The interior Dirichlet problem $\Delta^2 u = 0$ in Ω, $\nabla u = \mathbf{g}\in C_p$ on $\partial\Omega$ is solvable with $u = u_m((I+\mathscr{L})^{-1}\mathbf{g}; X)$.* The exterior Dirichlet problem $\Delta^2 u = 0$ *in $\bar{\Omega}^C$, $\nabla u = \mathbf{g}\in C_p$ on $\partial\Omega$ is solvable with $u = u_m((-I+\mathscr{L})^{-1}\mathbf{g}_0; X) + \nabla w$ where for $\mathbf{g}\in C_p$, $\mathbf{g}$ can be written uniquely as $\mathbf{g} = \mathbf{g}_0 + \nabla w$ with $\mathbf{g}_0\in(-I+\mathscr{L})(C_p)$ and $w = \alpha x + \beta y + \gamma(x^2+y^2) + \delta$.*

THEOREM 5.17. *For $\boldsymbol{\varphi}\in(L^q\times L^q)_0(\partial\Omega)$ we have the following*:

(i) *$\mathscr{L}^*\boldsymbol{\varphi}(Q)$ exists almost everywhere with respect to arclength, $\mathscr{L}$ is bounded from $(L^q\times L^q)_0(\partial\Omega)$ to itself in the $(L^q\times L^q)(\partial\Omega)$ norm and is compact from $(L^q\times L^q)_0(\partial\Omega)$ to itself.*

(ii) *The nontangential*

$$\lim_{X\to Q} \mathbf{L}v_m(X) = \begin{cases} (-I+\mathscr{L}^*)\boldsymbol{\varphi}(Q), & X\in\Omega, \\ (I+\mathscr{L}^*)\boldsymbol{\varphi}(Q), & X\notin\bar{\Omega} \end{cases}$$

for almost every $Q\in\partial\Omega$.

(iii) *$(-I+\mathscr{L}^*)^{-1}$ exists on $(L^q\times L^q)_0(\partial\Omega)$.*

COROLLARY 5.18. *The adjoint boundary value problem $\Delta^2 v = 0$ in Ω, $\mathbf{L}v = \boldsymbol{\varphi}\in(L^q\times L^q)_0(\partial\Omega)$ is solvable by $v = v_m((-I+\mathscr{L}^*)^{-1}\boldsymbol{\varphi}; X)$.*

Remark. It is important to note that the operator $-I+\mathscr{L}^*$ is not exactly the adjoint of $-I+\mathscr{L}$. The adjoint of $-I+\mathscr{L}$ acts on the dual of C_p, which is a coset space. The space $(L^q\times L^q)_0(\partial\Omega)$ is a function space that is close to the dual of C_p, however, work is required to extend the invertibility of the adjoint of $-I+\mathscr{L}$ from the dual of C_p to invertibility on $(L^q\times L^q)_0(\partial\Omega)$. The details of the proofs can be found in Cohen and Gosselin [7] and [8].

6. Cauchy formulae. In the theory of Hardy spaces of the upper half plane, functions $f\in L^p(\mathbb{R})$, $1<p<\infty$ can be identified with analytic functions $f(z)$ on the upper half plane satisfying $\sup_{y>0}\int |f(x+iy)|_{dx}^p < \infty$. This identification is obtained by convolving the boundary function f with the complex kernel $(i\pi z)^{-1}$. For an arbitrary C^1 (or even a Lipschitz domain if $f\in L^p(\partial\Omega)$ for $p\geqq 2$) we can obtain the same type of identification by applying the properties of the classical double layer potential to the Cauchy integral of the boundary data. (See Fabes, Jodeit, and Riviere [9], Fabes and Kenig [10], or Verchota [13] for more details.)

An analogous kind of identification of compatible triples of boundary functions with analytic functions of order two can be obtained from Burgatti's formula (2.11). (By compatible triples we mean the space $\mathscr{B}_p = \{\dot{f} = (f, g, h)\in L_1^p\times L^p\times L^p(\partial\Omega)$: $f_s = gx_s + hy_s$ almost everywhere with respect to arclength$\}$.) For $\dot{f} = (f, g, h)\in\mathscr{B}_p$ we define the complex potential

$$\varphi_{\dot{f}}(w) = \frac{1}{2\pi i}\int_{\partial\Omega}\frac{f(z)}{z-w}\,dz - \frac{1}{4\pi i}\int_{\partial\Omega}(g(z)+ih(z))\frac{\bar{z}-\bar{w}}{z-w}\,dz. \tag{6.1}$$

It follows from the article by Cohen and Gosselin on the Dirichlet problem for the biharmonic equation [6] that if $\mathcal{U}_f = \operatorname{Re} \varphi_f$ and $\dot{\mathcal{U}}_f = (\mathcal{U}_f, (\mathcal{U}_p)_x, (\mathcal{U}_p)_y)$, then there exists an invertible operator $T: \mathcal{B}_p \to \mathcal{B}_p$ such that the nontangential $\lim_{X \to P \in \partial\Omega} \dot{\mathcal{U}}_{T^{-1}\dot{f}}(X) = \dot{f}(P)$ almost everywhere. The map $\dot{f} \to \varphi_f$ then identifies the boundary space $\mathcal{B}_p$ with a space of functions analytic of order two in Ω.

We have seen in §§ 3 and 4 that the solutions to the traction problem and stationary Stokes problem can be obtained by solving a biharmonic problem and finding a biharmonic conjugate. An examination of the matrix kernels (5.8) and (5.10) suggests that we can obtain biharmonic conjugates to the solutions of (1.1) and (1.2) from a Cauchy-type formula if we can find a biharmonic conjugate to the function $\tilde{F}$.

But $\tilde{F}(z) = (-1/4\pi) \operatorname{Im}\{\bar{z}z \log z - \frac{1}{2}\bar{z}z + \frac{1}{2}z^2\}$ and $\bar{\partial}^2\{\bar{z}z \log z - \frac{1}{2}\bar{z}z + \frac{1}{2}z\} = 0$. If we let $F(z) = \operatorname{Re}(-1/4\pi)\{\bar{z}z \log z - \frac{1}{2}\bar{z}z + \frac{1}{2}z^2\}$, then $\bar{\partial}^2(F + i\tilde{F}) = 0 = i\bar{\partial}^2(\tilde{F} - iF)$. This implies that $\tilde{F} - iF$ is analytic of order two, which suggests the following Cauchy-type potentials.

DEFINITION 6.2. For $\mathbf{g} \in C_p$ define the complex modified multiple layer potential

$$(u_m + i\tilde{u}_m)(\mathbf{g}; X) = \int_{\partial\Omega} \mathbf{g}(Q)\mathbf{L}^Q(\tilde{F} - iF)(X - Q)^T \, ds(Q). \tag{6.3}$$

DEFINITION 6.4. For $\boldsymbol{\varphi} \in (L^q \times L^q)_0(\partial\Omega)$ define the complex modified lower-order potential

$$(v_m + i\tilde{v}_m)(\boldsymbol{\varphi}, X) = \int_{\partial\Omega} \varphi(\beta)(\tilde{F} - iF)_x(P - X) + \psi(P)(\tilde{F} - iF)(P - X) \, ds(P). \tag{6.5}$$

It then follows immediately that $(u_m + i\tilde{u}_m)((I + \mathcal{L})^{-1}(f_2, -f_1); X)$ gives the function analytic of order two, which solves the stationary Stokes problem and $(v_m + i\tilde{v}_m) \cdot ((-I + \mathcal{L}^*)^{-1}(-Y_n, X_n); X)$ solves the traction problem. That is, all dependent variables can be obtained from the real and imaginary parts of these complex potentials by differentiation. Furthermore, explicit integral representations of the stresses, displacements, velocity components, and pressure can be obtained from taking the appropriate derivatives of the matrix kernels.

REFERENCES

[1] S. AGMON, *Multiple layer potentials and the Dirichlet problem for higher order elliptic equations in the plane*, Commun. Pure Appl. Math., 10 (1957), pp. 179-230.

[2] AIRY, Brit., Assoc. Rep., 1862.

[3] P. BURGATTI, *Sulle funzioni analitiche d'ordine n*, Boll. Un. Mat. Ital., 1922.

[4] ———, *Sulle funzioni analitiche d'ordine n e sull' equilibria elastico in due dimensioni*, Boll. Un. Mat. Ital., 1923.

[5] A. P. CALDERON, *Cauchy integrals on Lipschitz curves and related operators*, Proc. Nat. Acad. Sci. U.S.A., 75, (1977), pp. 1324-1327.

[6] J. COHEN AND J. GOSSELIN, *The Dirichlet problem for the biharmonic equation in a bounded C^1 domain in the plane*, Indiana Math. J., 32 (1983), pp. 635-685.

[7] ———, *Adjoint boundary value problems for the biharmonic equations on C^1 domains in the plane*, Ark. Mat., 23 (1985), pp. 217-240.

[8] ———, *Stress potentials on C^1 domains*, J. Math. Anal. Appl., to appear.

[9] E. FABES, M. JODEIT, AND N. RIVIERE, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., (1979).

[10] E. FABES AND C. KENIG, *On the Hardy space of a C^1 domain*, Ark. Mat., 19 (1981), pp. 1-22.

[11] S. G. MIKHLIN, *Integral Equations*, Pergamon Press, 1957.

[12] N. MUSKHELISHVILI, *Some basic problems of the mathematical theory of elasticity*, P. Noordhoff, Groningen, the Netherlands, 1963.

[13] 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.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

008

BOUNDS FOR EFFECTIVE COEFFICIENTS OF PERIODIC FIBER-REINFORCED MATERIALS*

GELU I. PAŞA†

Abstract. This paper contains a generalization of the estimates obtained by Tartar [*Estimation de coefficients homogenisés*, Lecture Notes in Mathematics 704, Springer-Verlag, Berlin, New York, 1977, pp. 364–377] for linear elasticity, without the symmetry condition $A_{ijkh} = A_{khij}$. A problem of this type appears when dealing with nonorthogonal coordinates. Moreover, the paper includes a comparison to similar estimates obtained by Hashin and Rosen [3].

Key words. composite periodic materials, homogenizations, compensated compactness

AMS(MOS) subject classifications. 73K20, 35B40, 35B10

Introduction. We consider an elastic medium and assume that the associated coefficients A_{ijkh} are Y-periodic, Y being a parallelepiped in R^n. An example of such a medium may be given by the composite materials, i.e., a low-resistance matrix with periodically inserted fibers of higher resistance.

Among the composite materials, we are mostly interested in those for which the dimension of the periodicity cell Y is very small compared to that of the entire medium. By introducing a suitable topology, we can obtain convergence of the initially periodic and bounded coefficients, thus replacing the former problem with another one with constant coefficients, which may be determined by solving it for a single cell.

In order to handle the one-cell problem with better results, it is very helpful to have some a priori estimates for these constant coefficients (the so-called "homogenized" or "effective" coefficients). Among the methods already proposed in this direction, we refer in the following sections to variational methods, as well as to the homogenization method (see [1], [5]).

Using variational principles in linear elasticity, Hashin and Shtrikman in 1963 obtained limits for effective moduli of quasi-homogeneous and quasi-isotropic composite materials of arbitrary phase geometry [8].

In their 1964 paper, Hashin and Rosen studied a composite material with hexagonal periodicity and nonisotropic components [3]. Their method is based on the minimum principle for the potential energy, and they need an exact solution for the given configuration.

Estimates for the effective thermic conductivity of a composite material with general periodic configuration were obtained by Tartar in his 1976 paper [6], by using the homogenization method along with his own and F. Murat's method of compensated compactness (see [4]).

Tartar's results were generalized to the case of linear elasticity by Francfort and Murat in 1986 [7]. They deal with symmetric matrices of elasticity, and the composite materials have isotropic components. Their paper also includes a comparison of their results with those from [8]. While the "bulk modulus" limits are the same in both papers, the "shear modulus" limits are more accurate in [8].

Here we present a different type of generalization for Tartar's results, namely, the case of linear elasticity without the symmetry condition $A_{ijkh} = A_{khij}$. Such a situation

* Received by the editors May 27, 1986; accepted for publication (in revised form) February 25, 1988.
† Department of Mathematics, The National Institute for Scientific and Technical Creation, Badul Păcii 220, 79622 Bucharest, Romania.

may appear when applying Tartar's method to a configuration for which periodicity directions do not coincide with the coordinate axes, i.e., as in [3]. Moreover, we consider a two-phase composite with nonisotropic elastic components and compare our results with those obtained in [3]. We emphasize that instead of the symmetry condition we introduce condition (*) (in 1), which allows for an estimation of the upper bound limits for the matrix A (in the sense of Tartar). On the contrary, the lower bound limit cannot be obtained unless we deal with a particular element, which, in the configuration mentioned in [3], is of the "shear modulus" type. The second section contains an example as well as a comparison between our results and the estimations obtained by Hashin and Rosen [3]; their limits are more accurate, in accordance with the result mentioned in [7].

1. Bounds for homogenized coefficients. As we have already mentioned in the Introduction, the principal results of this section (contained in Proposition 2 and Theorem 2) give an estimation of the homogenized coefficients appearing in linear elasticity, obtained by using a generalization of Tartar's method along with an idea of McConnel [2].

The following definitions will be needed in the sequel.

DEFINITION 1. Let $A_{ijkh} = A_{ijhk} = A_{jikh}$ be Y-periodic and strongly elliptic:

$$A^{\varepsilon}_{ijkh}(x) = A_{ijkh}(x/\varepsilon).$$

Let the equation

$$-\partial(A^{\varepsilon}_{ijkh}\partial u^{\varepsilon}_k/\partial x_h)/\partial x_i = F_j \quad \text{in } \Omega \subset R^n,$$
$$\underline{u}^{\varepsilon}/\partial\Omega = 0.$$

Using the result of the homogenization theory (Sanchez-Palencia [5], Benssoussan, Lions, and Papanicolaou [1]) it is possible to prove the existence of A^0_{ijkh} and $\underline{u}^0$ such that $\underline{u}^{\varepsilon} \rightharpoonup \underline{u}^0$ in the weak topology of the Sobolev space $H^1_0(\Omega)$, when $\varepsilon \to 0$, and

$$-\partial(A^0_{ijkh}\partial u^0_k/\partial x_h)/\partial x_i = F_j \quad \text{in } \Omega \subset R^n,$$
$$\underline{u}^0/\partial\Omega = 0.$$

In this case, A^{ε}_{ijkh} are H-convergent to A^0_{ijkh}:

$$A^{\varepsilon}_{ijkh} \overset{H}{\rightharpoonup} A^0_{ijkh}.$$

DEFINITION 2. Assume M and N are two "matrices" of coefficients such that $M_{ijkh} = M_{ijhk} = M_{jikh}$. Then $M \leqq N \Leftrightarrow (Mc, c) \leqq (Nc, c)$ for all $c_{ij} = c_{ji} \in R^3 \times R^3$, $(,)$ being the scalar product in R^9.

Remark 1. If $A^{\varepsilon}_{ijkh} \overset{H}{\rightharpoonup} A^0_{ijkh}$, then $A^{\varepsilon}_{ijkh}(\partial u^{\varepsilon}_k/\partial x_h)(\partial u^{\varepsilon}_i/\partial x_j) \rightharpoonup A^0_{ijkh}(\partial u^0_k/\partial x_h) \times (\partial u^0_i/\partial x_j)$ weakly in $L_2(\Omega)$ as a direct consequence of the convergence theorem. Notice that in the local problem we use the transposed matrix of A^{ε}_{ijkh}.

PROPOSITION 1. *Let $p^{\varepsilon}_{ij} \in H^1(\Omega)$ be such that $p^{\varepsilon}_{ij} \rightharpoonup p^*_{ij}$ weakly in $L_2(\Omega)$ and $\|\partial(p^{\varepsilon}_{ij})/\partial x_i\|_{L_2(\Omega)} \leqq$ const. Assume that $\underline{u}^{\varepsilon} \in H^1(\Omega)$ are such that $\underline{u}^{\varepsilon} \rightharpoonup u^*$ weakly in $H^1(\Omega)$. Then*

$$\int_{\Omega} p^{\varepsilon}_{ij}\partial u^{\varepsilon}_i/\partial x_j\phi \to \int_{\Omega} p^*_{ij}\partial u^*_i/\partial x_j\phi \quad \forall \phi \in C^{\infty}_0(\Omega).$$

Proof. Define

$$\underline{h}^{\varepsilon_i} = (p^{\varepsilon}_{i1}, p^{\varepsilon}_{i2}, \cdots, p^{\varepsilon}_{in}), \quad \underline{v}^{\varepsilon_i} = (\partial u^{\varepsilon}_i/\partial x_1, \partial u^{\varepsilon}_i/\partial x_2, \cdots, \partial u^{\varepsilon}_i/\partial x_n).$$

Then we have $\|\operatorname{div}(\underline{h}^{\varepsilon_i})\|_{L_2(\Omega)} \leqq$ const,

$$(\operatorname{rot} \underline{v}^{\varepsilon_i})_{sq} = \partial^2 u^{\varepsilon}_i/\partial x_s\partial x_q - \partial^2 u^{\varepsilon}_i/\partial x_q\partial x_s = 0,$$

and we may apply the compensated compactness method for vectors, (cf. [1]; see also [4]). □

Let us consider now the following assumption concerning the elastic coefficients:

(∃) A coefficient with $j = kh$ (for example, A_{1313}) such that

(*) $A_{13sq} = A_{sq13}$.

First we shall prove a result that does not use the compensated compactness method.

THEOREM 1. *Let* $A^\varepsilon \overset{H}{\rightharpoonup} Q$ *and* $A^\varepsilon \rightharpoonup A$, $(A^\varepsilon)^{-1} \rightharpoonup (B)^{-1}$ *weakly in* $L_2(\Omega)$. *Then*

(a) $Q \geqq A$,

(b) $Q_{1313} \geqq B_{1313}$.

Proof. (a) We consider $c_{ij} = c_{ji} \in R^3 \times R^3$ and start with

$$(A^\varepsilon(v^\varepsilon - c), v^\varepsilon - c) \geqq 0, \tag{1}$$

where v^ε is defined by

$$\begin{aligned} &\partial(A^\varepsilon_{ijkh}\partial u^\varepsilon_k/\partial x_h)/\partial x_i = 0, \qquad v^\varepsilon_{kh} = \partial u^\varepsilon_k/\partial x_h, \\ &u^\varepsilon_k - c_{kh}\cdot x_h \in H^1_0(\Omega). \end{aligned} \tag{2}$$

Passing to the limit with $\varepsilon \to 0$ in (1) and using $v^\varepsilon \rightharpoonup c$ we obtain

$$(Qc, c) - 2\cdot(Qc, c) + (Ac, c) \geqq 0.$$

Consequently, $Q \geqq A$. Therefore, condition (*) was not necessary for the first part of the theorem.

(b) In order to determine the lower bound we consider

$$((A^\varepsilon)^{-1}(A^\varepsilon v^\varepsilon + d), A^\varepsilon v^\varepsilon + d) \geqq 0, \tag{3}$$

where v^ε is given by (2) for $c_{13} \neq 0$ and $c_{sq} = 0$, $sq \neq 13$, and $d_{ij} = -B_{ijkh}c_{kh}$. We have

$$(A^\varepsilon)^{-1}_{pqst}A^\varepsilon_{stij}\frac{\partial u^\varepsilon_i}{\partial x_j}A^\varepsilon_{pqmn}\frac{\partial u^\varepsilon_m}{\partial x_n} = \delta_{pi}\cdot\delta_{qj}A^\varepsilon_{pqmn}\frac{\partial u^\varepsilon_i}{\partial x_j}\frac{\partial u^\varepsilon_m}{\partial x_n} \rightharpoonup (Qc, c), \tag{4}$$

$$(A^\varepsilon)^{-1}_{pqst}A^\varepsilon_{stij}\partial u^\varepsilon_i/\partial x_j d_{pq} = \delta_{pi}\cdot\delta_{qj}\partial u^\varepsilon_i/\partial x_j \cdot d_{pq} \rightharpoonup (d, c), \tag{5}$$

$$(A^\varepsilon)^{-1}_{pqst}\cdot d_{pq}\cdot d_{st} \rightharpoonup (B^{-1}d, d), \tag{6}$$

the convergences being in the weak topology of $L_2(\Omega)$. Notice that only the symmetric part of $\partial u^\varepsilon_i/\partial x_j$ appeared in the above relations and that we have used the inversion formula $A_{ijkh}(A)^{-1}_{khpq} = \delta_{ip}\cdot\delta_{jq}$. A closer look at the three terms defined above shows that no condition of symmetry for A^ε was ever needed.

Let us consider, for any $\phi \in C^\infty_0(\Omega)$,

$$\begin{aligned} I &\equiv \int_\Omega (A^\varepsilon)^{-1}_{pqst}d_{st}A^\varepsilon_{pqij}\frac{\partial u^\varepsilon_i}{\partial x_j}\phi \\ &= \int_\Omega (A^\varepsilon)^{-1}_{pqst}d_{st}A^\varepsilon_{pqij}\left(\frac{\partial u^\varepsilon_i}{\partial x_j} - c_{ij}\right)\phi + \int_\Omega (A^\varepsilon)^{-1}_{pqst}d_{st}A^\varepsilon_{pqij}c_{ij}\phi \\ &\equiv I_1 + I_2. \end{aligned}$$

Since the "vector" c and the "matrices" A^ε, B, and $(A^\varepsilon)^{-1}$ are all bounded, it follows that

$$I_1 \leqq \text{const.}\left|\int_\Omega (\partial u^\varepsilon_i/\partial x_j - c_{ij})\phi\right| \to 0.$$

Condition (*) together with $c_{ij}=0$ for $ij\neq 13$ imply

$$I_2=\int_\Omega (A^\varepsilon)^{-1}_{pqst}d_{st}A^\varepsilon_{pq13}c_{13}\phi=\int_\Omega \delta_{s1}\delta_{t3}d_{st}c_{13}\phi=\int_\Omega d_{13}c_{13}\phi.$$

Consequently, we have proved that

$$(A^\varepsilon)^{-1}_{pqst}d_{st}A^\varepsilon_{pqij}\partial u^\varepsilon_i/\partial x_j \rightharpoonup d_{13}c_{13}, \tag{7}$$

weakly in $L_2(\Omega)$. Using (3)-(7) we obtain

$$Q_{1313}c_{13}^2+2\cdot d_{13}c_{13}+(B)^{-1}_{pqst}d_{pq}d_{st}\geqq 0. \tag{8}$$

But

$$\begin{aligned}(B)^{-1}_{pqst}d_{pq}d_{st}&=(B)^{-1}_{pqst}d_{pq}(-B_{stmn}c_{mn})\\&=-(B)^{-1}_{pqst}B_{st13}c_{13}d_{pq}=-d_{13}c_{13}.\end{aligned}$$

Notice that any symmetry properties of A^ε are inherited by B and A, since they are obtained by averaging the coefficients of A^ε. In the sample presented in § 2 we will prove the condition (*) for A^ε. The last relation together with relation (8) implies

$$Q_{1313}c_{13}^2+d_{13}c_{13}\geqq 0,\qquad Q_{1313}c_{13}^2-B_{1313}c_{13}^2\geqq 0.$$

Therefore, condition (*) was needed only in the last part of the theorem.

In [6] Tartar showed that the bounds obtained by using the above theorem are not precise enough when A^ε is not continuous and the values of A^ε in the fiber tend to zero or to infinity.

To obtain better limits, we define the following notion of convergence, which generalizes that introduced by Tartar.

DEFINITION 3. $A^\varepsilon \overset{13}{\rightharpoonup} A$ if and only if

$$\begin{aligned}&1/A^\varepsilon_{1313}\rightharpoonup 1/A_{1313},\qquad A^\varepsilon_{13sq}/A^\varepsilon_{1313}\rightharpoonup A_{13sq}/A_{1313},\\&A^\varepsilon_{sq13}/A^\varepsilon_{1313}\rightharpoonup A_{sq13}/A_{1313},\\&A^\varepsilon_{sqkh}-A^\varepsilon_{sq13}A^\varepsilon_{13kh}/A^\varepsilon_{1313}\rightharpoonup A_{sqkh}-A_{sq13}A_{13kh}/A_{1313},\end{aligned}$$

where all convergences are in the weak topology of $L_2(\Omega)$ and $(sq),(kh)\neq(13)$.

First we prove a preliminary result.

PROPOSITION 2. *Let $A^\varepsilon \overset{13}{\rightharpoonup} A$ and $c_{kh}=c_{hk}\in R^3\times R^3$. Assume $A^\varepsilon_{13sq}=A^\varepsilon_{sq13}$. Then there exists $w^\varepsilon_{kh}\in L_2(\Omega)$ such that we have the following*:

(1) $w^\varepsilon_{kh}\rightharpoonup c_{kh}$ *weakly in* $L_2(\Omega)$.

(2) $A^\varepsilon_{ijkh}w^\varepsilon_{ij}w^\varepsilon_{kh}\rightharpoonup A_{ijkh}c_{ij}c_{kh}$ *weakly in* $L_2(\Omega)$.

Proof. Define

$$\begin{aligned}&p^\varepsilon_{ij}=A^\varepsilon_{ijkh}w^\varepsilon_{kh}\quad\text{for } (ij)\neq(13),\\&p^\varepsilon_{13=K}\ \text{(a constant that will be defined later)},\\&w^\varepsilon_{13}=\text{variable},\qquad w^\varepsilon_{sq}=c_{sq}\quad\text{for } (sq)\neq(13).\end{aligned}$$

For $(sq)\neq(13)$ we have

$$\begin{aligned}&K=p^\varepsilon_{13}=A^\varepsilon_{1313}w^\varepsilon_{13}+A^\varepsilon_{13sq}c_{sq},\\&w^\varepsilon_{13}=K/A^\varepsilon_{1313}-A^\varepsilon_{13sq}c_{sq}/A^\varepsilon_{1313}\rightharpoonup K/A_{1313}-A_{13sq}c_{sq}/A_{1313}.\end{aligned}$$

By replacing w^ε_{13} we obtain, for (ij) and $(sq)\neq(13)$

$$p^\varepsilon_{ij}\rightharpoonup c_{sq}(A_{ijsq}-A_{ij13}A_{13sq}/A_{1313})+K\cdot A_{ij13}/A_{1313}.$$

Using the last relation, we can compute the weak limit of the product $p^\varepsilon_{rt} w^\varepsilon_{rt}$, since p^ε_{13} and w^ε_{sq} are constant for $(sq) \neq (13)$:

$$\begin{aligned} p^\varepsilon_{rt} w^\varepsilon_{rt} \rightharpoonup\ & K(K/A_{1313} - A_{13sq}c_{sq}/A_{1313}) \\ & + c_{ij}\{A_{ij13}K/A_{1313} + c_{sq}(A_{ijsq} - A_{ij13}A_{13sq}/A_{1313})\} \\ = & K^2/A_{1313} - K \cdot A_{13sq}c_{sq}/A_{1313} + K \cdot A_{ij13}c_{ij}/A_{1313} \\ & + c_{sq}A_{ijsq}c_{ij} - c_{ij}c_{sq}A_{ij13}A_{13sq}/A_{1313}. \end{aligned}$$

From the conditions $A^\varepsilon_{13sq} = A^\varepsilon_{sq13}$ and $K = c_{13}A_{1313} + A_{13sq}c_{sq}$ it follows that

$$p^\varepsilon_{rt} w^\varepsilon_{rt} \rightharpoonup A_{1313}c^2_{13} + 2 \cdot c_{13}A_{13sq}c_{sq} + c_{sq}c_{ij}A_{sqij}.$$

Taking into account the choice of K, we obtain $w^\varepsilon_{13} \rightharpoonup c_{13}$. □

Now we define the following:

MA3 = arithmetical mean with respect to $x_1, x_2, x_4, \cdots, x_n$,

MA2 = arithmetical mean with respect to $x_1, x_3, x_4, \cdots, x_n$.

THEOREM 2. *Let* MA3 $(A^\varepsilon) \xrightarrow{13} B$, $A^\varepsilon \xrightarrow{H} Q$, $A^\varepsilon_{13sq} = A^\varepsilon_{sq13}$. *Then* $Q \leqq B$.

Proof. We consider $c_{ij} = c_{ji} \in R^3 \times R^3$, w^ε_{ij} the sequence defined in Proposition 2 for MA3 (A^ε) and c_{ij}, and v^ε_{kh} given by (2). Then we have the following convergence relations in the weak sense of $L_2(\Omega)$:

$$\begin{aligned} & v^\varepsilon_{kh} \rightharpoonup c_{kh}, \qquad w^\varepsilon_{ij} \rightharpoonup c_{ij}, \\ & \text{MA3}\,(A^\varepsilon v^\varepsilon, v^\varepsilon) \rightharpoonup \text{MA3}\,(Qc, c) = (Qc, c), \\ & \text{MA3}\,(A^\varepsilon v^\varepsilon, w^\varepsilon) \rightharpoonup \text{MA3}(Qc, c) = (Qc, c), \\ & \text{MA3}\,(A^\varepsilon w^\varepsilon, w^\varepsilon) = (\text{MA3}\,(A^\varepsilon) w^\varepsilon, w^\varepsilon) \rightharpoonup (Bc, c), \end{aligned}$$

where we have used the fact that, for $(ij) \neq (13)$, w^ε_{ij} are constant and w^ε_{13} depends only on x_3; therefore

$$\begin{aligned} & \text{rot}\,(0, 0, w^\varepsilon_{13}, 0, \cdots, 0) = 0, \\ & \|\text{div}\,(p^\varepsilon_{11}, p^\varepsilon_{12}, p^\varepsilon_{13}, \cdots p^\varepsilon_{1n})\|_{L_2(\Omega)} \leqq \text{const}, \end{aligned}$$

with $p^\varepsilon_{ij} = A^\varepsilon_{ijkh} \partial u^\varepsilon_k / \partial x_h$. Now we may use the compensated compactness method for MA3 $(A^\varepsilon v^2, w^\varepsilon)$. Starting with

$$\text{MA3}\,(A^\varepsilon(v^\varepsilon - w^\varepsilon), v^\varepsilon - w^\varepsilon) \geqq 0, \tag{10}$$

we obtain, by using the above convergence relations,

$$(Qc, c) - 2(Qc, c) + (Bc, c) \geqq 0.$$

THEOREM 3. *Let* A^ε *satisfy the condition* (*), MA2 $[(A^\varepsilon)^{-1}] \xrightarrow{13} (E)^{-1}$, $A^\varepsilon \xrightarrow{H} Q$. *Then* $Q_{1313} \geqq E_{1313}$.

Proof. We consider v^ε_{kh} given by (2) with $c_{13} \neq 0$, $c_{sq} = 0$ for $(sq) \neq (13)$, $d_{ij} = -E_{ijkh}c_{kh}$, and w^ε_{kh} the sequence given in Proposition 2 for MA2 $[(A^\varepsilon)^{-1}]$ and the constants d_{ij}. Considering

$$\text{MA2}\,((A^\varepsilon)^{-1}(A^\varepsilon v^\varepsilon + w^\varepsilon), A^\varepsilon v^\varepsilon + w^\varepsilon) \geqq 0, \tag{11}$$

we notice that w^ε_{13} depends only on x_2. Consequently,

$$\begin{aligned} & \text{div}\,(w^\varepsilon_{13}, 0, \cdots, 0) = 0, \qquad & & \text{rot}\,(\partial u^\varepsilon_3/\partial x_1, \partial u^\varepsilon_3/\partial x_2, \cdots, \partial u^\varepsilon_3/\partial x_n) = 0, \\ & \text{div}\,(0, 0, w^\varepsilon_{13}, 0, \cdots, 0) = 0, \qquad & & \text{rot}\,(\partial u^\varepsilon_1/\partial x_1, \partial u^\varepsilon_1/\partial x_2, \cdots, \partial u^\varepsilon_1/\partial x_n) = 0. \end{aligned}$$

Using the compensated compactness method for MA2$(v^\varepsilon, w^\varepsilon)$, we obtain

$$\int_\Omega w_{13}^\varepsilon(\partial u_1^\varepsilon/\partial x_3+\partial u_3^\varepsilon/\partial x_1)/2\cdot\phi\to\int_\Omega c_{13}d_{13}\phi,\quad \forall\phi\in C_0^\infty(\Omega).$$

Remark 1, together with condition (*), implies the following convergence relations, in the weak sense of $L_2(\Omega)$:

$$MA2\,((A^\varepsilon)^{-1}(A^\varepsilon v^\varepsilon), A^\varepsilon v^\varepsilon)\rightharpoonup(Qc, c),$$
$$\mathrm{MA2}((A^\varepsilon)^{-1}(A^\varepsilon v^\varepsilon), w^\varepsilon)\rightharpoonup d_{13}c_{13},$$
$$\mathrm{MA2}\,((A^\varepsilon)^{-1}(w^\varepsilon), w^\varepsilon)=(\mathrm{MA2}\,(A^\varepsilon)^{-1}w^\varepsilon, w^\varepsilon)\rightharpoonup(E^{-1}d, d).$$

Taking into account condition (*) we obtain, by a proof similar to the second part of Theorem 1, MA2 $((A^\varepsilon)^{-1}w^\varepsilon, A^\varepsilon v^\varepsilon)\rightharpoonup d_{13}c_{13}$. The above convergence relations together with relation (11) imply, as in Theorem 1, that $Q_{1313}-E_{1313}\geqq 0$. □

2. A particular configuration. The example given in this section is meant to illustrate the above theoretical results. Hashin and Rosen studied in [3] a composite material formed by cylindrical fibers of radii r, disposed in the vertex of a hexagonal array, surrounded by a matrix. Their method is based on an exact solution in cylindrical coordinates, and for this reason the interior of each hexagon is partially filled by the matrix, which is formed of cylinders of radii $\frac{1}{2}$ if the edge lengths of these hexagons are equal to one. We consider the medium formed by periodic parallelograms (Fig. 1).

The stress-strain relation is written in the reduced form:

$$\begin{aligned}
&\sigma_{11}=C_{11}e_{11}+C_{12}e_{22}+C_{12}e_{33},\\
&\sigma_{22}=C_{12}e_{11}+C_{22}e_{22}+C_{23}e_{33},\\
&\sigma_{33}=C_{12}e_{11}+C_{23}e_{22}+C_{22}e_{33},\\
&\sigma_{12}=2C_{44}e_{12},\qquad \sigma_{13}=2C_{44}e_{13},\\
&\sigma_{23}=(C_{22}-C_{23})\cdot e_{23},
\end{aligned}\tag{12}$$

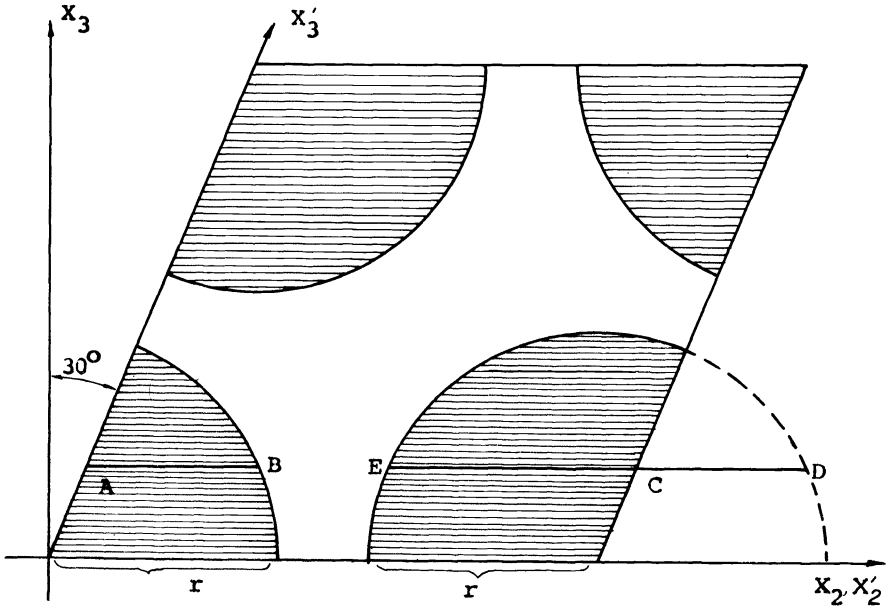


FIG. 1

where the usual six-by-six matrix notation has been used. C_{11} stands for a_{1111}, C_{12} stands for a_{1122}, and so on. The coefficients have the usual symmetry properties and depend only on x_2 and x_3; $a_{1313}=a_{1212}$.

The same type of isotropy is considered for the corresponding homogeneous medium. In [6] the estimates are obtained in a coordinate system parallel to the directions of a periodicity. We have to transform the original Cartesian form of the equation

$$\partial(a^{\varepsilon}_{ijkh}\partial u^{\varepsilon}_k/\partial x_h)/\partial x_j = F_i \tag{13}$$

for the new system since this one has angles of 60° between the coordinate axes. If $\underline{g}_k$ and $\underline{f}_p$ are, respectively, the unit vectors of the Cartesian and oblique coordinates, we have

$$\underline{g}_k = \alpha_{pk}\cdot \underline{f}_p, \qquad \underline{f}_p = \beta_{kp}\cdot \underline{g}_k,$$

$$\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1/\sqrt{3} \\ 0 & 0 & 2/\sqrt{3} \end{pmatrix}, \qquad \beta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & \sqrt{3}/2 \end{pmatrix}.$$

Then from (13) we obtain

$$\partial(A^{\varepsilon}_{pqst}\partial u^{\varepsilon'}_s/\partial x'_t)/\partial x'_q = F'_p$$

where $'$ denotes the components in the new system and

$$2\cdot A_{pqst} = \alpha_{pi}\beta_{qj}(\beta_{hs}\alpha_{tk}+\beta_{ks}\alpha_{th})a_{ijkh}.$$

Consequently, the stress-strain relations in the new system have the following form:

$$\begin{aligned}
&\sigma'_{11} = a_{1111}e'_{11}+a_{1122}e'_{22}+a_{1133}e'_{33},\\
&\sigma'_{22} = (4a_{2211}/3)e'_{11}+(5a_{2222}-a_{3322})/3\cdot e'_{22}+(5a_{2233}-a_{2222})/3\cdot e'_{33}-(4a_{2323}/3)e'_{23},\\
&\sigma'_{33} = (4a_{3311}/3)e'_{11}+(4a_{3322}/3)e'_{22}+(4a_{3333}/3)e'_{33},\\
&\sigma'_{12} = (4a_{1313}/3)e'_{12}-(2a_{1313}/3)e'_{13},\\
&\sigma'_{13} = -(2a_{1313}/3)e'_{12}+(4a_{1313}/3)e'_{13},\\
&\sigma'_{23} = -(2a_{3311}/3)e'_{11}-(2a_{2222}/3)e'_{22}-(2a_{2233}/3)e'_{33}+(4a_{2323}/3)e'_{23},
\end{aligned} \tag{14}$$

therefore $A_{1313}=A_{1212}=4a_{1313}/3$. Notice that the matrix A^{ε}_{pqst} satisfies the conditions (*): $A^{\varepsilon}_{13sq}=A^{\varepsilon}_{sq13}$.

We note that

$$(A^{\varepsilon})^{-1}_{12st}=(A^{\varepsilon})^{-1}_{st12} \quad \text{and} \quad (A\varepsilon)^{-1}_{13st}=(A^{\varepsilon})^{-1}_{st13}.$$

These are direct consequences of the fact that columns 12 and 13 become proportional after we eliminate row 13, as well as rows 12 and 13, which become proportional after we eliminate column 13. Therefore

$$(A^{\varepsilon})^{-1}_{sq13}=0, \quad (A^{\varepsilon})^{-1}_{sq12}=0, \quad (A^{\varepsilon})^{-1}_{13sq}=0, \quad (A^{\varepsilon})^{-1}_{12sq}=0$$

for $sq\neq 13$ and 12. The minors of the elements A^{ε}_{1312} and A^{ε}_{1213} are both equal to

$$-(2a^{\varepsilon}_{1313}/3)\begin{vmatrix} A^{\varepsilon}_{1111} & A^{\varepsilon}_{1122} & A^{\varepsilon}_{1133} & 0 \\ A^{\varepsilon}_{2211} & A^{\varepsilon}_{2222} & A^{\varepsilon}_{2233} & A^{\varepsilon}_{2323} \\ A^{\varepsilon}_{2311} & A^{\varepsilon}_{2322} & A^{\varepsilon}_{2323} & 0 \end{vmatrix},$$

and, consequently, we have $(A^{\varepsilon})^{-1}_{13sq}=(A^{\varepsilon})^{-1}_{sq13}$, $(A^{\varepsilon})^{-1}_{12sq}=(A^{\varepsilon})^{-1}_{sq12}$.

The element $(A^\varepsilon)^{-1}_{1313}$, obtained using the development of $\det(A^\varepsilon)$ with respect to the row A^ε_{12sq}, is given by

$$(A^\varepsilon)^{-1}_{1313} = 1/a^\varepsilon_{1313}. \tag{15}$$

We note that $(A^\varepsilon)^{-1}_{1313}$ is used to construct the sequence w^ε defined in Theorem 3; it is important to see that the homogenized coefficient corresponding to a^ε_{1313} is estimated in [3] using only the values of this coefficient; in the method presented here, the estimates may generally depend on the other coefficients.

Following Definition 3, we must compute the 13-limit of the harmonical mean with respect to x'_3 and the 13-limit of the arithmetical mean with respect to x'_2 of the coefficient A^ε_{1313}, thus obtaining upper and lower bounds for it:

$$\lim_{13\ \varepsilon\to 0} \mathrm{MA3}\,(A^\varepsilon_{1313}) = B_{1313}, \qquad \lim_{13\ \varepsilon\to 0} \mathrm{MA2}\,[(A^\varepsilon)^{-1}_{1313}] = E^{-1}_{1313}.$$

Finally we obtain $E_{1313} \leqq Q_{1313} \leqq B_{1313}$, where $A^\varepsilon \xrightarrow{H} Q$. Using the method described here, it is possible to obtain bounds for the effective coefficients in the case when the space between the periodical fibers is completely filled by the matrix of low resistance. We consider this situation in the sequel.

Now we analyze the geometry of the periodicity cell (Fig. 1). In the vertex of the romb of edge lengths 1 and 60° in the origin are situated sectors of radii r and 60° (respectively, 120°), filled by the fiber. The rest of the romb is filled by the matrix of low resistance. We can see that $AB = CD$. For simplicity we put $x'_3 = t$, $x'_2 = s$. We have (see Fig. 2)

$$t = OM, \qquad OZ = t\sqrt{3}/2,$$
$$ZH = (\sqrt{4r^2 - 3t^2})/2, \qquad UH = 2 \cdot ZH = \sqrt{4r^2 - 3t^2}.$$

The maximum value of t for which a parallel line to Os intersects the circle of radii r is $ON = 2r/\sqrt{3}$. We consider the case when such a parallel may intersect only two circles (the centers of which are situated on Os), then we consider $r \leqslant \sqrt{3}/4$, which is in accordance with the above hypothesis.

We let $A_{1313} = h$ in the fiber, $A_{1313} = 1$ in the matrix.

For the upper bound we have

$$\mathrm{MA2}\,(A_{1313}) = \begin{cases} h\sqrt{4r^2 - 3t^2} + 1 - \sqrt{4r^2 - 3t^2}, & t \in (0, 2r/\sqrt{3}), \\ 1, & t \in (2r/\sqrt{3}, 1/2), \end{cases}$$

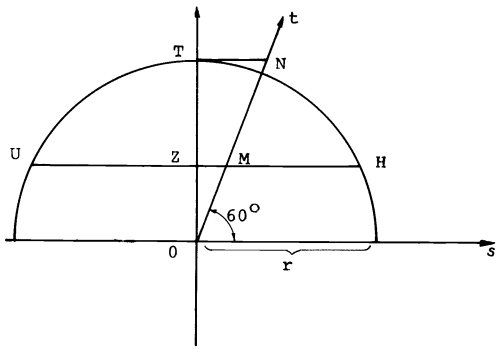


FIG. 2

and using the fact that the weak limit in $L_2(\Omega)$ of a periodic function of (x/ε), for $\varepsilon \to 0$, is the mean value on a cell (see [5]), we obtain

$$A^+_{1313} = \left\{ 2 \cdot \int_0^{2r/\sqrt{3}} [(h-1)\sqrt{4r^2-3t^2}+1]^{-1}\, dt + 1 - 4r/\sqrt{3} \right\}^{-1}. \tag{17}$$

The lower bound is given by

$$A^-_{1313} = 2 \cdot \int_0^{2r/\sqrt{3}} \left[\left(\frac{1}{h} - 1 \right) \sqrt{4r^2-3t^2}+1 \right]^{-1} dt + 1 - 4r/\sqrt{3}. \tag{18}$$

Hashin and Rosen [3] obtained the following bounds:

$$G^+(m_g v_1 + v_2), \tag{19}$$

$$G^-(v_1/m_g + v_2)^{-1} \tag{20}$$

where

$$m_g = [h(1+b^2)+1-b^2]/[h(1-b^2)+1+b^2], \qquad b = 2r,$$

$$v_1 = 0.918, \qquad v_2 = 0.082 = 1 - v_1.$$

In order to compare these two methods, we compute the expressions (17)-(20) for different values of h and r. Considering, for example, $r = 0.2$ and $r = 0.3$, we obtain the results in Tables 1 and 2. We can see that in general the estimates given in [3] are more precise than those given by the method presented here.

TABLE 1
$r = 0.2$.

h	A^-_{1313}	G^-	G^+	A^+_{1313}
2	0.819	1.103	1.112	0.840
3	0.846	1.157	1.159	0.907
6	0.879	1.232	1.236	1.037
10	0.889	1.275	1.282	1.140
20	0.908	1.302	1.310	1.236
30	0.909	1.314	1.323	1.271
50	0.912	1.323	1.333	1.315
100	0.914	1.331	1.341	1.351
⋮	⋮	⋮	⋮	⋮
∞	0.917	1.339	1.349	1.393

TABLE 2
$r = 0.3$.

h	A^-_{1313}	G^-	G^+	A^+_{1313}
5	1.086	1.551	1.579	1.341
15	1.198	1.785	1.844	1.846
35	1.236	1.872	1.945	2.116
65	1.250	1.905	1.984	2.249
115	1.258	1.922	2.005	2.326
195	1.262	1.931	2.016	2.371
⋮	⋮	⋮	⋮	⋮
∞	1.268	1.945	2.032	2.441

We want to emphasize that only for A_{1313} is it possible to obtain both upper and lower limits using the above theoretical results. Generally, only the upper limit may be computed, since A^ε verifies the condition (*) (see Definition 2).

Hashin and Rosen [3] also presented the bounds corresponding to the following constants:

$$K_{23} = (a_{2222} + a_{2233}), \qquad G_{23} = (a_{2222} - a_{2233}),$$
$$E_1 = a_{1111} - 2a_{1122}/(a_{2222} + a_{2233}).$$

If we consider $\lambda_{22} = \lambda_{33} = 1$, $\lambda_{ij} = 0$ for $(ij) \neq (22)$ and (33), then relations (14) and (12) give us

$$A_{ijkh}\lambda_{ij}\lambda_{kh} = 8 \cdot (a_{2222} + a_{2233})/3.$$

For $\lambda_{23} = 1$, $\lambda_{ij} = 0$ if $(ij) \neq (23)$, we obtain

$$A_{ijkh}\lambda_{ij}\lambda_{kh} = 4 \cdot a_{2323}/3 = 4(C_{22} - C_{23})/3$$
$$= 4(a_{2222} - a_{2233})/3.$$

Therefore we can apply our method to obtain upper bounds for the constants K_{23} and G_{23}. Since in general the coefficients with $ij \neq kh$ cannot be directly estimated, it seems that the method presented here is not applicable for E_1 (which represents the longitudinal Young modulus).

REFERENCES

[1] A. BENSSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[2] W. H. MCCONNEL, *On the approximation of elliptic operators with discontinuous coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 34 (1976), pp. 121-137.

[3] Z. HASHIN AND W. ROSEN, *The elastic moduli of fiber-reinforced materials*, J. Appl. Mech. Trans. ASME, Ser. E, 31 (1964), pp. 223-232.

[4] F. MURAT, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 36 (1978), pp. 489-507.

[5] E. SANCHEZ-PALENCIA, *Non-Homogeneous Media and Vibration Theory*, Lecture Notes in Physics 127, Springer-Verlag, Berlin, New York, 1980.

[6] L. TARTAR, *Estimation de coefficients homogenisés*, Lecture Notes in Mathematics 704, Springer-Verlag, Berlin, New York, 1977, pp. 364-377.

[7] G. A. FRANCFORT AND F. MURAT, *Homogenization and optimal bounds in linear elasticity*, Arch. Rational Mech. Anal., 94 (1986), pp. 307-335.

[8] Z. HASHIN AND S. SHTRIKMAN, *A variational approach to the theory of the elastic behaviour of multiphase materials*, J. Mech. Phys. Solids, 11 (1963), pp. 127-140.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

009

SMOOTHING PROPERTIES OF LINEAR VOLTERRA INTEGRODIFFERENTIAL EQUATIONS*

W. DESCH† AND R. GRIMMER‡

Abstract. Boundary value problems for hyperbolic linear partial differential integral equations of convolution type on an interval are studied. A necessary and sufficient condition on the convolution kernel is given such that discontinuities of the boundary data are smoothed in the interior of the interval. The result is applied to consider dynamics of viscoelastic media.

Key words. integrodifferential equations, Rayleigh problem, viscoelastic material, propagation of singularities

AMS(MOS) subject classifications. 45K05, 73D99

1. Introduction. We study propagation of singularities by a linear first-order hyperbolic partial differential integral equation in one space variable, namely

$$u_t(t,\xi) = au_\xi(t,\xi) + cu(t,\xi) + \int_0^t [d(t-s)u_\xi(s,\xi) + \dot{h}(t-s)u(s,\xi)]\,ds, \tag{1}$$
$$(t \geqq 0, \quad \xi \in (\xi_1, \xi_2))$$

with initial boundary conditions

$$u(0,\xi) = 0, \quad b_j u(t,\xi_j) = v_j(t) \quad (j = 1, 2).$$

u is considered to be vector valued and a, c, d, h, b_j are matrices. Thus the problem includes the standard first-order transcription of

$$u_{tt}(t,\xi) = G(0)u_{\xi\xi}(t,\xi) + \int_0^t G'(t-s)u_{\xi\xi}(s,\xi)\,ds \tag{2}$$

with suitable initial and boundary conditions.

The latter equation describes the dynamics of a linearly viscoelastic homogeneous medium (given sufficient symmetry to reduce the problem to one space dimension). (See, e.g., [2], [3].) Considerable effort has been spent investigating how a discontinuity in the initial or boundary data is propagated to the interior of the interval. A major part of the literature, such as [1], [6]-[9], [12]-[15], [20a-c], treats the case that the kernel d is sufficiently smooth (i.e., it admits at least two locally integrable derivatives). Reference [5] considers a similar situation in a nonlinear setting. In this case, a jump discontinuity in the boundary data gives rise to a jump discontinuity in the solution, traveling back and forth between the boundaries. The presence of a memory term does not affect the hyperbolic character of the equation, not even the wave speed of the discontinuity; its only effect is exponential damping of the stepsize of the discontinuity. However, in [19], Renardy has pointed out that the scene changes drastically as d is allowed to be unbounded at $t = 0$. He shows that singular kernels may generate solutions

* Received by the editors June 8, 1987; accepted for publication (in revised form) April 25, 1988.

† Institut für Mathematik, Universität Graz, Brandhofgasse 18, A-8010 Graz, Austria. Part of this work was done while this author was at Southern Illinois University, Carbondale, Illinois 62901. The Austrian Fonds zur Förderung der Wissenschaftlichen Forschung provided travel money by the Austrian-American Cooperative Research Program grant P5691.

‡ Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901. This author's research was partially supported by National Science Foundation grant DMS-8701552.

that are infinitely often differentiable in the interior of the interval, though discontinuous on the boundary. This fits very well to a result in [10], showing that unboundedness of the kernel together with suitable monotonicity conditions implies that the resolvent operator to the integrodifferential equation is compact; smoothing is likely to be expected in this case.

A recent paper by Hrusa and Renardy [11] shows that the situation is even a bit more complicated. When kernels with different types of singularities at $t=0$ are compared, they indicate that whether the solution is C^∞ across the characteristics at any time $t>0$ or does not gain any differentiability at all depends on the boundedness of $G'(t)/\ln(t)$ rather than $G'(t)$. In the limiting case that G' has a logarithmic singularity, the solution gains smoothness gradually as time increases. (Dealing with inhomogeneous boundary data rather than initial data, this means that the solution becomes smoother and smoother as we proceed deeper into the space interval.)

While writing this paper, we became aware of a contemporaneous study by Prüss, [17], which examines an integrated version of (2). The second derivative of u with respect to ξ is replaced by the generator of a cosine family, while the scalar function $G(t)$ may be replaced by a Stieltjes measure of specified type that includes the case G is completely monotone. We will compare our work with both [11] and [17] in the last section of this paper.

While the work in [11] is done by explicit computation of various examples, this paper means to give a general theorem characterizing the kernels d that lead to smoothing of the solution of a first-order hyperbolic system. Moreover, we give a formula for the space interval needed to gain one degree of differentiability in the limiting case. Unfortunately, our results for the general case (1) are very technical. If $d(t)$ is a scalar multiple $\varphi(t)d$ of a constant matrix d, considerable simplifications can be carried out. In particular, if φ is convex decreasing, smoothing depends on $\lim_{t\to 0}\varphi(t)/\ln(t)$, in accordance with [11].

In the last section of this paper, we apply our results to some nonscalar problems in viscoelasticity where we encounter equations in which $d(t)$ is not a scalar multiple of a constant matrix, e.g., a Timoshenko beam. It will be seen that many such problems can be treated with our techniques.

2. The main theorems. We consider the following equation:

$$
\begin{aligned}
&u_t(t,\xi) = au_\xi(t,\xi) + cu(t,\xi) + \int_0^t [d(t-s)u_\xi(s,\xi) + h(t-s)u(s,\xi)]\,ds,\\
&b_j u(t,\xi_j) = v_j(t) \qquad (j=1,2),\\
&u(0,\xi) = 0\\
&\text{for } t\geqq 0,\ \xi\in[\xi_1,\xi_2].
\end{aligned}
\tag{2.1}
$$

The solution u is supposed to be an $\mathbb{R}^n$-valued function on $[0,\infty)\times[\xi_1,\xi_2]$. Our hypotheses on the coefficients are the following.

Hypothesis 2.1. (a) a is a real $n\times n$-matrix that has n distinct real eigenvalues $\lambda_1<\cdots<\lambda_m<0<\lambda_{m+1}<\cdots<\lambda_n$. By $\{e_i\colon i=1\cdots n\}$ we denote a system of eigenvectors of a corresponding to the eigenvalues λ_i, respectively. $\{f_i\colon i=1\cdots n\}$ is a system of eigenvectors of the transposed matrix a^*, such that $f_i^* e_j=\delta_{ij}$. r is the $n\times n$-matrix $(f_1\cdots f_n)^*$; thus $r^{-1}=(e_1\cdots e_n)$, and $rar^{-1}=\operatorname{diag}(\lambda_1\cdots\lambda_n)$. We put $\kappa_i=\lambda_i^{-1}$, and for $i=1\cdots m$, $\delta_i=-1$, while $\delta_i=1$ for $i=m+1\cdots n$.

(b) Putting $m_1 = m$ and $m_2 = n - m$, we assume for $j = 1, 2$ that b_j is a real $m_j \times n$-matrix and that the matrices $(b_1 e_1 \cdots b_1 e_m)$ and $(b_2 e_{m+1} \cdots b_2 e_n)$ are invertible. (We may assume without loss of generality that they are $m_j \times m_j$ unit matrices.)

(c) c is a real $n \times n$-matrix.

(d) For some $\gamma \geqq 0$, the function $t \to e^{-\gamma t} d(t)$ is a real $n \times n$-matrix valued integrable function on $[0, \infty)$ and $te^{-\gamma t} d(t)$ is in $W^{1,1}(0, \infty)$.

(h) The function $t \to e^{-\gamma t} h(t)$ is a real $n \times n$-matrix valued integrable function on $[0, \infty)$.

(v) The functions $t \to e^{-\gamma t} v_j(t)$, $(j = 1, 2)$ are in $L^2([0, \infty), \mathbb{R}^{m_j})$.

(Hypothesis 2.1(a), (b), (c), (v) makes (2.1) without the convolution terms a hyperbolic first-order initial boundary value problem that is well posed in $L^2([\xi_1, \xi_2], \mathbb{R}^n)$. Parts (d), (h) of the hypothesis are minimal boundedness assumptions on the convolution kernels.)

Before we can state our second hypothesis, we need to introduce the Laplace transform $\hat{}$:

$$\hat{d}(\tau) = \int_0^\infty e^{-t\tau} d(t)\, dt \quad \text{(which exists for Re } \tau \geqq \gamma).$$

Since $\lim_{|\tau| \to \infty,\, \mathrm{Re}\, \tau \geqq \gamma} \hat{d}(\tau) = 0$, we may choose some θ sufficiently large such that for $\mathrm{Re}\, \tau \geqq \theta$ the matrix $a + \hat{d}(\tau)$ has n distinct eigenvalues $\lambda_1(\tau) \cdots \lambda_n(\tau)$ (converging to $\lambda_1 \cdots \lambda_n$ as $|\tau| \to \infty$), and that $\operatorname{sgn} \mathrm{Re}\, \lambda_i(\tau) = \operatorname{sgn} \lambda_i = \delta_i \neq 0$. In particular, $(a + \hat{d}(\tau))^{-1}$ exists with eigenvalues $\kappa_i(\tau) = \lambda_i^{-1}(\tau)$.

Hypothesis 2.2. There exists a constant M such that for $\rho \geqq \theta$, $\sigma \in \mathbb{R}$, $j = 1 \cdots n$, $\delta_j \sigma\, \mathrm{Im}\, \kappa_j(\rho + i\sigma) \leqq M$.

(A discussion of ways to verify this hypothesis will be given in § 4.)

Taking Laplace transforms in (2.1), we obtain

$$\begin{aligned} &\tau \hat{u}(\tau, \xi) = (a + \hat{d}(\tau)) \hat{u}_\xi(\tau, \xi) + (c + \hat{h}(\tau)) \hat{u}(\tau, \xi), \\ &b_j \hat{u}(\tau, \xi_j) = \hat{v}_j(\tau) \qquad (j = 1, 2) \\ &\text{for Re } \tau \text{ sufficiently large and } \xi \in [\xi_1, \xi_2]. \end{aligned} \tag{2.2}$$

It has been suggested by one of the referees that Hypothesis 2.2 may be equivalent to the well-posedness of (2.1). It seems clear that this hypothesis is intimately related to the well-posedness of (2.1) in $L^2((\xi_1, \xi_2), \mathbb{R}^n)$ with arbitrary boundary matrices b_j that are subject to Hypothesis 2.1; however, the proof of such a relation would seem to be quite lengthy and technical. Being interested in propagation of singularities rather than problems of well-posedness, we confine ourselves to the following existence result.

THEOREM 2.1. *Suppose Hypotheses* 2.1 *and* 2.2 *are valid. Then there exists a unique function* $u : [0, \infty) \times [\xi_1, \xi_2] \to \mathbb{R}^n$, *such that for sufficiently large* θ *and each* $\xi \in [\xi_1, \xi_2]$, *the function* $t \to e^{-\theta t} u(t, \xi)$ *is in* $L^2([0, \infty), \mathbb{R}^n)$, *and the Laplace transform* $\hat{u}(\tau)$ *satisfies* (2.2).

To state our main result on smoothing we introduce the numbers

$$\tilde{\beta}_1(\tau) = \max_{j = 1 \cdots m} (-\mathrm{Im}\, \kappa_j(\tau)), \qquad \tilde{\beta}_2(\tau) = \max_{j = m+1 \cdots n} (\mathrm{Im}\, \kappa_j(\tau))$$

for $\mathrm{Re}\, \tau \geqq \theta$, $\mathrm{Im}\, \tau \geqq 0$, and

$$\alpha_j^+ = \lim_{\sigma \to \infty} \sup -\sigma \tilde{\beta}_j(\theta + i\sigma)/\ln \sigma, \qquad \alpha_j^- = \lim_{\sigma \to \infty} \inf -\sigma \tilde{\beta}_j(\theta + i\sigma)/\ln \sigma.$$

(By Hypothesis 2.2, $-\sigma \tilde{\beta}_j(\theta + i\sigma)$ is bounded from below, thus $\alpha_j^\pm \in [0, \infty]$.)

THEOREM 2.2. *Suppose Hypotheses* 2.1 *and* 2.2 *are valid and choose* θ *sufficiently large (according to the hypotheses of this section and Theorem* 2.1*). Let* $k>0$ *be a real number. If* $(\xi-\xi_1)\alpha_1^- > k$ *and* $(\xi_2-\xi)\alpha_2^- > k$, *then for all boundary data* v_j *according to* (v) *the solution* $t \to e^{-\theta t}u(t,\xi)$ *is in* $W^{k,2}([0,\infty),\mathbb{R}^n)$. *If* $(\xi-\xi_1)\alpha_1^+ < k$ *or* $(\xi_2-\xi)\alpha_2^+ < k$, *then there exist boundary data* v_1 *and* v_2 *such that* $e^{-\theta t}u(t,\xi)$ *is not contained in* $W^{k,2}([0,\infty),\mathbb{R}^n)$.

Thus, if $\alpha_j^- = \alpha_j^+ = \alpha_j$, then $\alpha_1(\alpha_2)$ may be viewed as the reciprocal of the distance from the left (right) boundary needed to gain one degree of differentiability. If $\alpha_j^+ = 0$, no differentiability is obtained unless the boundary data are sufficiently smooth themselves. On the other hand, if $\alpha_j^- = \infty$, all boundary data are immediately smoothed to C^∞. Notice also, that $\alpha_j^\pm$ depends only on d and a, while c and h have no bearing on the degree of smoothing.

Remark. The definitions of the $\alpha_j^\pm$ are independent of θ for sufficiently large θ. This follows from the assumption that $e^{-\gamma t}t\,d(t) \in W^{1,1}(0,\infty)$. This independence is, in fact, the only reason for this smoothness assumption. To see this independence, we note that because a has distinct nonzero eigenvalues, we can appeal to the differentiability of $(a+\hat{d}(\tau))^{-1}$ to obtain $|\kappa_j(\tau_1)-\kappa_j(\tau)| = O(\|\hat{d}(\tau_1)-\hat{d}(\tau)\|)$ where $\tau_1 = \theta_1 + i\sigma$ and $\tau = \theta + i\sigma$. However, for $\theta > \theta_1$,

$$\hat{d}(\tau)-\hat{d}(\tau_1) = \int_0^\infty e^{-(\theta_1+i\sigma)t}\frac{[e^{-(\theta-\theta_1)t}-1]}{t}\,t\,d(t)\,dt$$
$$= O(\sigma^{-1}),$$

as it is the transform of a $W^{1,1}$ function. Hence,

$$\left|\frac{\sigma}{\ln\sigma}(\kappa_j(\tau_1)-\kappa_j(\tau))\right| \to 0 \quad \text{as } \sigma\to\infty.$$

3. Proof of Theorems 2.1 and 2.2. Throughout this section, let us assume that θ is sufficiently large, and $\tau = \rho + i\sigma$ with $\rho \geqq \theta$ and $\sigma > 0$.

From (2.2) we obtain

$$\hat{u}_\xi(\tau,\xi) = \tau(a+\hat{d}(\tau))^{-1}(I-\tau^{-1}(c+\hat{h}(\tau)))\hat{u}(\tau,\xi) =: \tau l(\tau)\hat{u}(\tau,\xi). \tag{3.1}$$

(I denotes the $n\times n$ unit matrix.)

As θ is chosen sufficiently large and $\lim_{|\tau|\to\infty,\,\mathrm{Re}\,\tau\geqq\gamma} l(\tau) = a^{-1}$, the matrix $l(\tau)$ possesses n distinct eigenvalues $\mu_1(\tau)\cdots\mu_n(\tau)$ converging to $\kappa_1\cdots\kappa_n$. In particular, $\operatorname{sgn}\mathrm{Re}\,\mu_j(\tau) = \delta_j$. Now, for $|\tau|\to\infty$, $\|l(\tau)-(a+\hat{d}(\tau))^{-1}\| = O(|\tau^{-1}|)$; thus also $|\mu_j(\tau)-\kappa_j(\tau)| = O(|\tau^{-1}|)$. This implies that Hypothesis 2.2 as well as the values for $\alpha_j^\pm$ remain the same if κ_i is replaced by μ_i and $\tilde{\beta}_j$ is replaced by $\beta_1 = \max_{j=1\cdots m} -\mathrm{Im}\,\mu_j(\tau)$, $\beta_2 = \max_{j=m+1\cdots n}\mathrm{Im}\,\mu_j(\tau)$.

Moreover, $l(\tau)$ can be diagonalized by a matrix $r(\tau)$ converging to r as $|\tau|\to\infty$, so that

$$r(\tau)l(\tau)r^{-1}(\tau) = \operatorname{diag}(\mu_1(\tau)\cdots\mu_n(\tau)) = k(\tau)\to\operatorname{diag}(\kappa_1\cdots\kappa_n).$$

We write $r^{-1}(\tau)$ in the form $(e_1(\tau)\cdots e_n(\tau))$ where $e_j(\tau)\to e_j$.

Putting $\tilde{u}(\tau,\xi) = r(\tau)\hat{u}(\tau,\xi)$, we obtain from (3.1) and the boundary conditions

$$\tilde{u}_\xi(\tau,\xi) = \tau k(\tau)\tilde{u}(\tau,\xi), \qquad b_j r^{-1}(\tau)\tilde{u}(\tau,\xi_j) = \hat{v}_j(\tau). \tag{3.2}$$

Consequently, the ith coefficient $\tilde{u}_i$ of $\tilde{u}$ satisfies

$$\tilde{u}_i(\tau,\xi) = \exp(\tau\mu_i(\tau)(\xi-\xi_1))\tilde{u}_i(\tau,\xi_1) = \exp(-\tau\mu_i(\tau)(\xi_2-\xi))\tilde{u}_i(\tau,\xi_2).$$

We put $U(\tau)=(\tilde{u}_1(\tau,\xi_1)\cdots\tilde{u}_m(\tau,\xi_1),\tilde{u}_{m+1}(\tau,\xi_2)\cdots\tilde{u}_n(\tau,\xi_2))^t$ and let U_i be the ith coefficient of U. So

$$\begin{aligned}&\tilde{u}_i(\tau,\xi)=\exp(-\delta_i\tau\mu_i(\tau)(\xi-\xi_1))U_i(\tau)\quad\text{for } i=1\cdots m,\quad\text{and}\\&\tilde{u}_i(\tau,\xi)=\exp(-\delta_i\tau\mu_i(\tau)(\xi_2-\xi))U_i(\tau)\quad\text{for } i=m+1\cdots n.\end{aligned}\tag{3.3}$$

To estimate the solution, we use Hypothesis 2.2 and obtain

$$\operatorname{Re}(\delta_i\tau\mu_i(\tau))=\rho\delta_i\operatorname{Re}\mu_i(\tau)-\sigma\delta_i\operatorname{Im}\mu_i(\tau)\geqq\theta\delta_i\operatorname{Re}\mu_i(\tau)-M.$$

As $\operatorname{Re}\mu_i(\tau)$ is bounded away from 0 (for large θ) and θ may be chosen large, we can achieve that $\operatorname{Re}(\delta_i\tau\mu_i(\tau))>0$ is arbitrarily large. In particular, by (3.3), we have $\|\tilde{u}(\tau,\xi)\|\leqq\|U(\tau)\|$.

The values of $U(\tau)$ are determined by the boundary conditions: As the matrices $b_1(e_1\cdots e_m)$ and $b_2(e_{m+1}\cdots e_n)$ are invertible, the matrices $p_1(\tau)=b_1(e_1(\tau)\cdots e_m(\tau))$ and $p_2(\tau)=b_2(e_{m+1}(\tau)\cdots e_n(\tau))$ are also (provided θ is chosen large). We put $q_1(\tau)=b_1(e_{m+1}(\tau)\cdots e_n(\tau))$ and $q_2(\tau)=b_2(e_1(\tau)\cdots e_m(\tau))$. Then the boundary conditions can be written in the form

$$\begin{bmatrix}\tilde{u}_1(\tau,\xi_1)\\\vdots\\\tilde{u}_m(\tau,\xi_1)\end{bmatrix}+p_1^{-1}(\tau)q_1(\tau)\begin{bmatrix}\tilde{u}_{m+1}(\tau,\xi_1)\\\vdots\\\tilde{u}_n(\tau,\xi_1)\end{bmatrix}=p_1^{-1}(\tau)\hat{v}_1(\tau),$$

$$\begin{bmatrix}\tilde{u}_{m+1}(\tau,\xi_2)\\\vdots\\\tilde{u}_n(\tau,\xi_2)\end{bmatrix}+p_2^{-1}(\tau)q_2(\tau)\begin{bmatrix}\tilde{u}_1(\tau,\xi_2)\\\vdots\\\tilde{u}_m(\tau,\xi_2)\end{bmatrix}=p_2^{-1}(\tau)\hat{v}_2(\tau).$$

Let $Q(\tau)$ be the block matrix

$$\begin{bmatrix}0 & p_2^{-1}(\tau)q_2(\tau)\operatorname{diag}_{i=1\cdots m}(e^{\tau\mu_i(\tau)(\xi_2-\xi_1)})\\ p_1^{-1}(\tau)q_1(\tau)\operatorname{diag}_{i=m+1\cdots n}(e^{-\tau\mu_i(\tau)(\xi_2-\xi_1)}) & 0\end{bmatrix}.$$

Then we obtain from (3.3)

$$U(\tau)+Q(\tau)U(\tau)=\begin{bmatrix}p_1^{-1}(\tau)\hat{v}_1(\tau)\\p_2^{-1}(\tau)\hat{v}_2(\tau)\end{bmatrix}.\tag{3.4}$$

Choosing θ sufficiently large, we can achieve that $\|Q(\tau)\|<1/4n$, so that there exists a unique solution $U(\tau)$ to (3.4) with some bound $\|U(\tau)\|\leqq N(\|\hat{v}_1(\tau)\|+\|\hat{v}_2(\tau)\|)$.

If $\theta\geqq\gamma$, the function $\sigma\to\hat{v}_j(\theta+i\sigma)$ is in $L^2(\mathbb{R},\mathbb{C}^n)$ by Plancherel's theorem. Hence for fixed $\xi\in[\xi_1,\xi_2]$, $\hat{u}(\theta+i\sigma,\xi)=r^{-1}(\theta+i\sigma)\tilde{u}(\theta+i\sigma,\xi)$ is also in $L^2(\mathbb{R},\mathbb{C}^n)$. Inverting the Laplace transform by a contour integral from $\theta-i\infty$ to $\theta+i\infty$, we obtain, again by Plancherel's theorem, that the function $t\to e^{-\theta t}u(t,\xi)$ is contained in $L^2(\mathbb{R},\mathbb{R}^n)$. This proves Theorem 2.1.

Let us now assume that $\alpha_1^-(\xi-\xi_1)>k$ and $\alpha_2^-(\xi_2-\xi)>k$, i.e., for sufficiently large $\sigma>0$,

$$-\sigma\beta_1(\theta+i\sigma)(\xi-\xi_1)/\ln\sigma>k\quad\text{and}\quad-\sigma\beta_2(\theta+i\sigma)(\xi_2-\xi)/\ln\sigma>k.$$

Consequently, for $j=1\cdots m$, $\sigma(\xi-\xi_1)\operatorname{Im}\mu_j(\theta+i\sigma)/\ln\sigma>k$, and for $j=m+1\cdots n$, $-\sigma(\xi_2-\xi)\operatorname{Im}\mu_j(\theta+i\sigma)/\ln\sigma>k$. Thus for sufficiently large σ and $j=1\cdots m$,

$$\begin{aligned}\operatorname{Re}((\theta+i\sigma)\mu_j(\theta+i\sigma)(\xi-\xi_1))&=\theta\operatorname{Re}\mu_j(\theta+i\sigma)(\xi-\xi_1)-\sigma\operatorname{Im}\mu_j(\theta+i\sigma)(\xi-\xi_1)\\&<-k\ln\sigma,\end{aligned}$$

while for $j = m+1 \cdots n$,

$$\operatorname{Re}((\theta+i\sigma)\mu_j(\theta+i\sigma)(\xi_2-\xi)) > k \ln \sigma.$$

By (3.3) we have for sufficiently large σ and $j = 1 \cdots m$,

$$\begin{aligned}|\tilde{u}_j(\theta+i\sigma, \xi)| &= \exp(\operatorname{Re}((\theta+i\sigma)\mu_j(\theta+i\sigma)(\xi-\xi_1)))|U_j(\theta+i\sigma)| \\ &\leqq \sigma^{-k}|U_j(\theta+i\sigma)|.\end{aligned}$$

Similarly for $j = m+1 \cdots n$,

$$\begin{aligned}|\tilde{u}_j(\theta+i\sigma, \xi)| &= \exp(-\operatorname{Re}((\theta+i\sigma)\mu_j(\theta+i\sigma)(\xi_2-\xi)))|U_j(\theta+i\sigma)| \\ &\leqq \sigma^{-k}|U_j(\theta+i\sigma)|.\end{aligned}$$

This implies that the function

$$\sigma \to \sigma^k \hat{u}(\theta+i\sigma, \xi) = \sigma^k r^{-1}(\theta+i\sigma)\tilde{u}(\theta+i\sigma, \xi)$$

is contained in $L^2([0,\infty), \mathbb{C}^n)$; thus by Plancherel's theorem, the function $t \to e^{-\theta t}u(t, \xi)$ is in $W^{k,2}([0,\infty), \mathbb{R}^n)$.

Assume conversely, that $(\xi-\xi_1)\alpha_1^+ < k$ or $(\xi_2-\xi)\alpha_2^+ < k$. Without loss of generality, we restrict our consideration to the first case; the second case is done similarly.

We choose some small $\varepsilon > 0$, such that $(\xi-\xi_1)\alpha_1^+ < k-2\varepsilon$, and a scalar function w such that $e^{-\gamma t}w(t) \in L^2([0,\infty), \mathbb{R})$, but $e^{-\theta t}w(t) \notin W^{\varepsilon,2}([0,\infty), \mathbb{R})$. Our boundary conditions are $v_1(t) = w(t) \cdot (1 \cdots 1)^t$, $v_2(t) = 0$. For sufficiently large $|\tau|$, we infer from (3.4) and $\|Q(\tau)\| \leqq 1/4n$ the following:

$$\left\| U(\tau) - \begin{bmatrix} p_1(\tau)\hat{v}_1(\tau) \\ p_2(\tau)\hat{v}_2(\tau) \end{bmatrix} \right\| \leqq \frac{1}{4n} \cdot \|U(\tau)\|; \quad \text{thus}$$

$$\|U(\tau)\| \leqq \frac{4}{3} \cdot \left\| \begin{bmatrix} p_1(\tau)\hat{v}_1(\tau) \\ p_2(\tau)\hat{v}_2(\tau) \end{bmatrix} \right\|, \quad \text{and}$$

$$\left\| U(\tau) - \begin{bmatrix} p_1(\tau)\hat{v}_1(\tau) \\ p_2(\tau)\hat{v}_2(\tau) \end{bmatrix} \right\| \leqq \frac{1}{3} \cdot \left\| \begin{bmatrix} p_1(\tau)\hat{v}_1(\tau) \\ p_2(\tau)\hat{v}_2(\tau) \end{bmatrix} \right\|.$$

Moreover, as $p_1(\tau) = b_1(e_1(\tau) \cdots e_m(\tau)) \to b_1(e_1 \cdots e_m) = I$, we have for sufficiently large $|\tau|$ that

$$p_1(\tau)(1 \cdots 1)^t = (\pi_1(\tau) \cdots \pi_m(\tau))^t \quad \text{with} \frac{3}{4} < |\pi_j(\tau)| < \frac{5}{4}.$$

Thus for $j = 1 \cdots m$

$$\begin{aligned}|U_j(\tau) - \hat{w}(\tau)\pi_j(\tau)| &\leqq 1/3n \cdot \|\hat{w}(\tau)(\pi_1(\tau) \cdots \pi_m(\tau), 0 \cdots 0)^t\| \\ &< 5|\hat{w}(\tau)|/12,\end{aligned}$$

$$\begin{aligned}|U_j(\tau)| &> |\hat{w}(\tau)\pi_j(\tau)| - 5|\hat{w}(\tau)|/12 > \left(\frac{3}{4} - \frac{5}{12}\right)|\hat{w}(\tau)| \\ &= |\hat{w}(\tau)|/3.\end{aligned}$$

As $(\xi-\xi_1)\alpha_1^+ < k-2\varepsilon$, we infer for sufficiently large σ that $-(\xi-\xi_1)\sigma\beta_1(\theta+i\sigma)/\ln\sigma < k-2\varepsilon$; thus there exists at least one $j \in \{1 \cdots m\}$ with $(\xi-\xi_1)\sigma \operatorname{Im}\mu_j(\theta+i\sigma) < (k-2\varepsilon)\ln\sigma$. Again considering sufficiently large σ, we obtain

$$\begin{aligned}(\xi-\xi_1)\operatorname{Re}((\theta+i\sigma)\mu_j(\theta+i\sigma)) &= (\xi-\xi_1)\theta \operatorname{Re}\mu_j(\theta+i\sigma) - (\xi-\xi_1)\sigma \operatorname{Im}\mu_j(\theta+i\sigma) \\ &> (\varepsilon-k)\ln\sigma.\end{aligned}$$

Thus

$$|\tilde{u}_j(\theta+i\sigma,\xi)| = \exp((\xi-\xi_1)\operatorname{Re}((\theta+i\sigma)\mu_j(\theta+i\sigma)))\cdot|U_j(\theta+i\sigma)|$$
$$> \sigma^{\varepsilon-k}|\hat{w}(\theta+i\sigma)|/3.$$

Suppose now that the function $t\to e^{-\theta t}u(t,\xi)$ is in $W^{k,2}([0,\infty),\mathbb{R}^n)$. By Plancherel's theorem, this implies that the function $\sigma\to\sigma^k\hat{u}(\theta+i\sigma,\xi)$; hence also $\sigma^k\tilde{u}(\theta+i\sigma,\xi)$ is in $L^2(\mathbb{R},\mathbb{C}^n)$. This implies $\sigma^\varepsilon\hat{w}(\theta+i\sigma)\in L^2(\mathbb{R},\mathbb{C})$, in contradiction to the assumption that $e^{-\theta t}w(t)\notin W^{\varepsilon,2}([0,\infty),\mathbb{R})$.

4. Simplifications and examples. The verification of Hypothesis 2.2, as well as the computation of the numbers $\alpha_j^\pm$, will in general give rise to tedious calculations, if at all possible. Therefore it seems worthwhile to point out some simplifications that may be applied to certain special cases to obtain explicit results.

We have already noticed in § 2, that the matrix c and the convolution kernel h do not contribute to the degree of smoothing. We improve this result and also show that bounded variation parts of the kernel d can be ignored.

PROPOSITION 4.1. *Suppose Hypothesis* 2.1 *is verified with two different kernels* d *and* $\tilde{d}$. *Moreover, let the function* $t\to e^{-\gamma t}(d(t)-\tilde{d}(t))$ *be of bounded variation and integrable on* $[0,\infty)$. *Then Hypothesis* 2.2 *is valid with* d *if and only if it is valid with* $\tilde{d}$, *and the values of* $\alpha_1^+(\alpha_1^-,\alpha_2^+,\alpha_2^-)$ *are the same for* d *and* $\tilde{d}$.

Proof. As a has distinct and nonzero eigenvalues, the eigenvalues and inverse matrices depend continuously differentiably on matrices in a sufficiently small neighborhood V of a. We pick θ sufficiently large, such that for $\operatorname{Re}\tau>\theta$ both matrices, $(a+\hat{d}(\tau))$ and $(a+\hat{\tilde{d}}(\tau))$, are contained in V. Thus the jth eigenvalue $\kappa_j(\tau)$ of $(a+\hat{d}(\tau))^{-1}$ and $\tilde{\kappa}_j(\tau)$ of $(a+\hat{\tilde{d}}(\tau))^{-1}$ satisfy $|\kappa_j(\tau)-\tilde{\kappa}_j(\tau)|=O(\|\hat{d}(\tau)-\hat{\tilde{d}}(\tau)\|)$. Since $e^{-\gamma t}(d(t)-\tilde{d}(t))$ is of bounded variation, $\|\hat{d}(\tau)-\hat{\tilde{d}}(\tau)\|=O(|\tau|^{-1})$. Thus $\tau(\kappa_j(\tau)-\tilde{\kappa}_j(\tau))$ is bounded for $\operatorname{Re}\tau>\theta$, implying that Hypothesis 2.2 is equivalent for the two cases. Moreover, $\sigma(\kappa_j(\theta+i\sigma)-\tilde{\kappa}_j(\theta+i\sigma))/\ln\sigma$ converges to 0 as $|\sigma|\to\infty$; thus we obtain the same values for $\alpha_1^+(\alpha_2^+,\alpha_1^-,\alpha_2^-)$ in both cases.

As an immediate consequence of this proposition and Theorem 2.2 we obtain Corollary 4.2.

COROLLARY 4.2. *Let Hypothesis* 2.1 *be satisfied with* d *such that* $e^{-\gamma t}d(t)$ *is integrable and of bounded variation. Then Hypothesis* 2.2 *is satisfied, and* L^2 *boundary data are not smoothed inside the interval.*

Proof. By Proposition 4.1 we may replace d by 0, so that $\kappa_j(\tau)=\kappa_j$ is real. This implies Hypothesis 2.2 and $\alpha_j^\pm=0$.

When we linearize the eigenvalues of $a+\hat{d}(\tau)$ at a, this leads to the following lemma.

LEMMA 4.3. *Suppose Hypothesis* 2.1 *holds. Then for* $|\tau|\to\infty$,

$$\operatorname{Im}\kappa_j(\tau) = -\lambda_j^{-2}\operatorname{Im}(f_j^*\hat{d}(\tau)e_j)+o(\|\operatorname{Im}\hat{d}(\tau)\|)+o(|\operatorname{Im}\kappa_j(\tau)|).$$

In particular, if $\|\operatorname{Im}\hat{d}(\tau)\|=O(|\operatorname{Im}(f_j^*\hat{d}(\tau)e_j)|)$, *then*

$$\operatorname{Im}\kappa_j(\tau)/\operatorname{Im}(f_j^*\hat{d}(\tau)e_j)\to-\lambda_j^{-2}.$$

Proof. $\operatorname{Im}(\lambda_j^2\kappa_j(\tau)|)=(\lambda_j^2-|\lambda_j^2(\tau)|)\operatorname{Im}\kappa_j(\tau)+|\lambda_j^2(\tau)|\operatorname{Im}\lambda_j^{-1}(\tau)=o(|\operatorname{Im}\kappa_j(\tau)|)-\operatorname{Im}\lambda_j(\tau)$. Since the eigenvalues of a are pairwise distinct, the eigenvalues $\lambda_j(q)$ of $a+q$ depend continuously differentiably on q, provided that q is a sufficiently small complex $n\times n$-matrix. Moreover, the derivative $\nabla\lambda_j(0)$ at $q=0$ is a linear map from the space

of $n \times n$-matrices into $\mathbb{C}$ given by $(\nabla\lambda_j(0))q = f_j^* q e_j$. If $|\tau|$ is large, so that $\hat{d}(\tau)$ is sufficiently small, then

$$\begin{aligned}\operatorname{Im}\lambda_j(\tau) &= \operatorname{Im}(\lambda_j(\tau)-\lambda_j)\\ &= \operatorname{Im}[i\nabla\lambda_j(\operatorname{Re}\hat{d}(\tau))\cdot\operatorname{Im}\hat{d}(\tau)+o(\|\operatorname{Im}\hat{d}(\tau)\|)]\\ &= \operatorname{Im}[i\nabla\lambda_j(0)\cdot\operatorname{Im}\hat{d}(\tau)+i(\nabla\lambda_j(\operatorname{Re}\hat{d}(\tau))-\nabla\lambda_j(0))\cdot\operatorname{Im}\hat{d}(\tau)]\\ &\quad+o(\|\operatorname{Im}\hat{d}(\tau)\|)\\ &= f_j^*\operatorname{Im}\hat{d}(\tau)e_j+o(\|\operatorname{Im}\hat{d}(\tau)\|).\end{aligned}$$

The hypothesis of this lemma again is technical enough; however, it is true if d is a scalar multiple of a constant matrix. The rest of this section is devoted to this case.

PROPOSITION 4.4. *Suppose that Hypothesis* 2.1 *is satisfied, and that* $d(t)=\varphi(t)d$ *with a scalar-valued function* φ *and a constant real matrix* d. *Moreover assume that for all* $j=1\cdots n$, $f_j^* de_j \neq 0$. *Then Hypothesis* 2.2 *is equivalent to the following. If* $\sigma\operatorname{Im}\hat{\varphi}(\rho+i\sigma)$ *is unbounded from above (below) for* $\rho\geqq\theta$, $\sigma\geqq 0$, *then for each* $j=1\cdots n$, $\delta_j f_j^* de_j > 0\ (<0)$. *Moreover, in this case we have*

$$\alpha_1^+ = \lim_{\sigma\to\infty}\sup[\sigma|\operatorname{Im}\hat{\varphi}(\theta+i\sigma)|/\ln\sigma]\cdot\min_{j=1\cdots m}\lambda_j^{-2}|f_j^* de_j|,$$

$$\alpha_1^- = \lim_{\sigma\to\infty}\inf[\sigma|\operatorname{Im}\hat{\varphi}(\theta+i\sigma)|/\ln\sigma]\cdot\min_{j=1\cdots m}\lambda_j^{-2}|f_j^* de_j|,$$

$$\alpha_2^+ = \lim_{\sigma\to\infty}\sup[\sigma|\operatorname{Im}\hat{\varphi}(\theta+i\sigma)|/\ln\sigma]\cdot\min_{j=m+1\cdots n}\lambda_j^{-2}|f_j^* de_j|,$$

$$\alpha_2^- = \lim_{\sigma\to\infty}\inf[\sigma|\operatorname{Im}\hat{\varphi}(\theta+i\sigma)|/\ln\sigma]\cdot\min_{j=m+1\cdots n}\lambda_j^{-2}|f_j^* de_j|.$$

Proof. Evidently, $\|\operatorname{Im}\hat{d}(\tau)\| = |\operatorname{Im}\hat{\varphi}(\tau)|\cdot\|d\| = O(|\operatorname{Im}\hat{\varphi}(\tau)f_j^* de_j|) = O(|\operatorname{Im} f_j^*\hat{d}(\tau)e_j|)$, so that the previous lemma applies and $\operatorname{Im}\kappa_j(\theta+i\sigma)/\operatorname{Im}\hat{\varphi}(\theta+i\sigma)\to -\lambda_j^{-2}f_j^* de_j$. Thus $\delta_j\operatorname{Im}\kappa_j(\theta+i\sigma)$ is bounded from above if and only if either $\operatorname{Im}\hat{\varphi}(\theta+i\sigma)$ is bounded, or $\operatorname{Im}\hat{\varphi}(\theta+i\sigma)$ is bounded from below, and $-\delta_j\lambda_j^{-2}f_j^* de_j<0$, or $\operatorname{Im}\hat{\varphi}(\theta+i\sigma)$ is bounded from above, and $-\delta_j\lambda_j^{-2}f_j^* de_j>0$. Now assume that one of these conditions is satisfied. Then

$$\begin{aligned}\tilde{\beta}_1(\theta+i\sigma) &= \max_{j=1\cdots m}(-\operatorname{Im}\kappa_j(\theta+i\sigma))\\ &= \max_{j=1\cdots m}[\lambda_j^{-2}f_j^* de_j\operatorname{Im}\hat{\varphi}(\theta+i\sigma)]+o(|\operatorname{Im}\hat{\varphi}(\theta+i\sigma)|).\end{aligned}$$

Employing the definition of α_1^+, we obtain

$$\alpha_1^+ = \lim_{\sigma\to\infty}\sup[-\max_{j=1\cdots m}\lambda_j^{-2}f_j^* de_j\operatorname{Im}\hat{\varphi}(\theta+i\sigma)/\ln\sigma],$$

which equals 0 if $\sigma|\operatorname{Im}\hat{\varphi}(\theta+i\sigma)|$ is bounded. If for some sequence $\sigma_k\to\infty$, $|\sigma_k\operatorname{Im}\hat{\varphi}(\theta+i\sigma_k)|\to\infty$, then for large k, $\operatorname{Im}\hat{\varphi}(\theta+i\sigma_k)$ and $f_j^* de_j$ have the opposite sign; thus

$$\alpha_1^+ = \lim_{\sigma\to\infty}\sup[\sigma|\operatorname{Im}\hat{\varphi}(\theta+i\sigma)|/\ln\sigma]\cdot\min_{j=1\cdots m}\lambda_j^{-2}|f_j^* de_j|.$$

The formulae for the other $\alpha_j^{\pm}$ are proved similarly.

Our last result shows how this is related to the singularity of the kernel $\varphi(t)$ at $t=0$.

PROPOSITION 4.5. *Suppose that $d(t)=\varphi(t)\,d$ with a constant $n\times n$-matrix d and a scalar, nonnegative, nonincreasing, convex function $\varphi(t)\,e^{-\gamma t}$. Moreover, for each $j=1\cdots n$ let $\delta_j f_j^*\,de_j<0$. Then Hypothesis 2.2 is satisfied, and*

$$\liminf_{t\to 0}[-\varphi(t)/\ln t]\cdot\min_{j=1\cdots m}|f_j^*\,de_j|\cdot\lambda_j^{-2}$$
$$\leqq\alpha_1^-\leqq\alpha_1^+\leqq 2\limsup_{t\to 0}[-\varphi(t)/\ln t]\cdot\min_{j=1\cdots m}|f_j^*\,de_j|\cdot\lambda_j^{-2},\quad \textit{and}$$
$$\liminf_{t\to 0}[-\varphi(t)/\ln t]\cdot\min_{j=m+1\cdots m}|f_j^*\,de_j|\cdot\lambda_j^{-2}$$
$$\leqq\alpha_2^-\leqq\alpha_2^+\leqq 2\limsup_{t\to 0}[-\varphi(t)/\ln t]\cdot\min_{j=m+1\cdots m}|f_j^*\,de_j|\cdot\lambda_j^{-2}.$$

In particular, if φ has a singularity weaker than logarithmic at 0, there are L^2 boundary data that never get smoothed in the interior of the interval. On the other hand, if φ has a singularity stronger than logarithmic, all boundary data are smoothed immediately to C^∞.

Proof. As for $\theta>\gamma$, the function $e^{-\gamma t}\varphi(t)$ is again nonnegative, convex, and nonincreasing; we may assume without loss of generality that φ itself is also, and put $\theta=0$. Since φ is nonincreasing, we obtain for $\sigma>0$, $\operatorname{Im}\hat\varphi(i\sigma)=-\int_0^\infty\sin(\sigma t)\varphi(t)\,dt<0$. Thus $\sigma\operatorname{Im}\hat\varphi(i\sigma)$ is bounded from above, and as $\delta_j f_j^*\,de_j<0$, Hypothesis 2.2 is satisfied.

To estimate the $\alpha_j^\pm$, we start out with

$$-\sigma\operatorname{Im}\hat\varphi(i\sigma)=\sigma\int_0^\infty\sin(\sigma t)\varphi(t)\,dt=\int_0^\infty\sin\tau\varphi\left(\frac{\tau}{\sigma}\right)d\tau$$
$$=\int_0^\pi\sin\tau\sum_{j=0}^\infty\left[\varphi\left(\frac{\tau+2\pi j}{\sigma}\right)-\varphi\left(\frac{\tau+\pi(2j+1)}{\sigma}\right)\right]d\tau.$$

Using the convexity of φ, we obtain

$$-\sigma\operatorname{Im}\hat\varphi(i\sigma)\geqq\int_0^\pi\sin\tau\sum_{j=0}^\infty\left[\varphi\left(\frac{\tau+2\pi j}{\sigma}\right)-\frac{1}{2}\left(\varphi\left(\frac{\tau+2\pi j}{\sigma}\right)+\varphi\left(\frac{\tau+2\pi(j+1)}{\sigma}\right)\right)\right]d\tau$$
$$=\frac{1}{2}\int_0^\pi\sin\tau\varphi\left(\frac{\tau}{\sigma}\right)d\tau,$$
$$-\sigma\operatorname{Im}\hat\varphi(i\sigma)=\int_0^\pi\sin\tau\left[\varphi\left(\frac{\tau}{\sigma}\right)-\sum_{j=1}^\infty\left[\varphi\left(\frac{\tau+\pi(2j-1)}{\sigma}\right)-\varphi\left(\frac{\tau+2\pi j}{\sigma}\right)\right]\right]d\tau$$
$$\leqq\int_0^\pi\sin\tau\varphi\left(\frac{\tau}{\sigma}\right)d\tau-\frac{1}{2}\int_0^\pi\sin\tau\sum_{j=1}^\infty\left[\varphi\left(\frac{\tau+\pi(2j-1)}{\sigma}\right)\right.$$
$$\left.-\varphi\left(\frac{\tau+\pi(2j+1)}{\sigma}\right)\right]d\tau$$
$$=\int_0^\pi\sin\tau\varphi\left(\frac{\tau}{\sigma}\right)d\tau-\frac{1}{2}\int_0^\pi\sin\tau\varphi\left(\frac{\tau+\pi}{\sigma}\right)d\tau\leqq\int_0^\pi\sin\tau\varphi\left(\frac{\tau}{\sigma}\right)d\tau.$$

Therefore,

$$\liminf_{\sigma\to\infty}(-\sigma\operatorname{Im}\hat\varphi(i\sigma)/\ln\sigma)\geqq\liminf_{\sigma\to\infty}\frac{1}{2\ln\sigma}\int_0^\pi\sin\tau\varphi\left(\frac{\tau}{\sigma}\right)d\tau$$
$$=\liminf_{\sigma\to\infty}\int_0^\pi\frac{\sin\tau(\ln\sigma-\ln\tau)}{2\ln\sigma}\cdot\frac{-\sigma(\tau/\sigma)}{\ln(\tau/\sigma)}\,d\tau$$
$$\geqq\liminf_{t\to 0}-\varphi(t)/\ln t\quad\text{since}$$

$$\int_0^\pi \frac{\sin \tau(\ln \sigma - \ln \tau)}{2 \ln \sigma}\, d\tau \to 1 \quad \text{as } \sigma \to \infty.$$

Similarly,

$$\begin{aligned}\limsup_{\sigma\to\infty} (-\sigma \operatorname{Im} \hat{\varphi}(i\sigma)/\ln \sigma) &\leqq \limsup_{\sigma\to\infty} \frac{1}{\ln \sigma} \int_0^\pi \sin \tau \varphi\left(\frac{\tau}{\sigma}\right) d\tau \\ &= \limsup_{\sigma\to\infty} \int_0^\pi \frac{\sin \tau(\ln \sigma - \ln \tau)}{\ln \sigma} \cdot \frac{-\varphi(\tau/\sigma)}{\ln(\tau/\sigma)}\, d\tau \\ &\leqq 2 \limsup_{t\to 0} -\varphi(t)/\ln t.\end{aligned}$$

This and Proposition 4.4 yield the desired estimates.

Example 1. To give an example, let us show how our results apply to a shear flow in a linearly viscoelastic fluid bordered by two parallel (infinite) plates. We assume that the fluid is initially quiescent and that it is then perturbed by moving the plates in a fixed, tangential direction. The resulting velocity field depends only on one space variable orthogonal to the boundaries. The direction of the velocity vector is parallel to the motion of the plates, so that the velocity field can be viewed as a scalar. Let $v(t, \xi)$ denote the velocity, $\sigma(t, \xi)$ the shear stress, and ρ the density of the fluid. $G(t)$ is the stress relaxation modulus. Then we have the constitutive equation

$$\sigma(t, \xi) = \int_0^t G(t-s) v_\xi(s, \xi)\, ds,$$

and the momentum equation

$$\rho v_t(t, \xi) = \sigma_\xi(t, \xi).$$

The initial and boundary conditions are

$$v(0, \xi) = 0, \quad \sigma(0, \xi) = 0, \quad v(t, \xi_1) = v_1(t), \quad v(t, \xi_2) = v_2(t).$$

Combining the equations and putting $u(t, \xi) = (v(t, \xi), \sigma(t, \xi))^t$, we obtain

$$\begin{aligned}u_t(t, \xi) &= \begin{bmatrix} 0 & \rho^{-1} \\ G(0) & 0 \end{bmatrix} u_\xi(t, \xi) + \int_0^t G'(t-s) \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} u_\xi(s, \xi)\, ds \\ &= a u_\xi(t, \xi) + \int_0^t G'(t-s)\, d u_\xi(s, \xi)\, ds,\end{aligned}$$

$$u(0, \xi) = 0, \quad b u(t, \xi_j) = (1, 0) u(t, \xi_j) = v_j(t) \quad (j = 1, 2).$$

It is a matter of elementary linear algebra to compute

$$\lambda_1 = -\left(\frac{G(0)}{\rho}\right)^{1/2}, \qquad \lambda_2 = \left(\frac{G(0)}{\rho}\right)^{1/2},$$

$$e_1 = \begin{bmatrix} 1 \\ -\sqrt{G(0)\rho} \end{bmatrix}, \qquad e_2 = \begin{bmatrix} 1 \\ \sqrt{G(0)\rho} \end{bmatrix},$$

$$f_1 = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2}\sqrt{G(0)\rho} \end{bmatrix}, \qquad f_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2}\sqrt{G(0)\rho} \end{bmatrix},$$

$$\delta_j f_j^*\, de_j = \tfrac{1}{2}\sqrt{G(0)\rho} \quad \text{for } j = 1, 2.$$

From physical considerations it is reasonable to assume that G' is negative, nondecreasing, convex down, and integrable. Replacing d by $-d$ and G' by $-G'$, we may apply Proposition 4.5 to see that Hypothesis 2.2 holds. Moreover, if G' has no

singularity or a singularity weaker than logarithmic at $t=0$, no smoothing occurs. If the singularity is stronger than logarithmic, the solutions are C^∞ with respect to time for each fixed ξ in the interior of the interval. In the limiting case that the singularity is logarithmic, the solution gains differentiability gradually as we go away from the boundary. The space interval needed to gain one degree of differentiability is α^{-1} with

$$\alpha = \sqrt{\rho}/2\sqrt{G(0)^3} \cdot \lim_{\sigma\to\infty} [\sigma \operatorname{Im} \hat{G}'(i\sigma)/\ln \sigma]$$

(provided that the limit exists, otherwise we must include the space interval between $(\alpha^+)^{-1}$ and $(\alpha^-)^{-1}$). This also implies that step discontinuities are smoothed immediately to continuous functions, as a step function is contained in $W^{1/2-\varepsilon,2}$ for any $\varepsilon>0$, while $W^{1/2+\varepsilon,2}$ is already embeddable in the space of continuous functions.

Let us briefly compare how our work relates to the results in [11]. The equation treated there is

$$u_{tt}(t,\xi) = \ell u_{\xi\xi}(t,\xi) + \int_{-\infty}^{t} m(t-s)(u_{\xi\xi}(t,\xi) - u_{\xi\xi}(s,\xi))\, ds.$$

Putting $u=v$, $\ell = \rho^{-1}(G(0) - \int_0^\infty G'(s)\, ds)$, $m(t) = -\rho^{-1}G'(t)$, this may be viewed as a second-order transcription of our problem. While we treat inhomogeneous boundary conditions on a finite interval, in [11] homogeneous boundary conditions and a nonzero initial condition are treated.

If the kernel m exhibits a singularity stronger than logarithmic, the boundary value problem shows immediate smoothing to C^∞. The same is true for the initial value problem, as far as smoothness across the characteristics is concerned. The "vertical" characteristic, showing always a singularity one degree weaker than the initial singularity, of course does not appear if the discontinuity comes in from the boundary. If the singularity of the kernel is weaker than logarithmic, no differentiability is gained in both cases. The effect observed in [11], that initial data of bounded variation may be smoothed to C^1, evidently requires more refined methods than the one developed in our paper.

The example of a kernel with logarithmic singularity in [11] is $m(t) = \sum_{k=0}^\infty \exp(-e^k t)$. It is shown that differentiability is gained gradually. It is easy to compute α by our formula:

$$\alpha = \sqrt{\rho}/2\sqrt{G(0)^3} \cdot \lim_{\sigma\to\infty} [\sigma \operatorname{Im} \hat{G}'(i\sigma)/\ln \sigma]$$

$$= \frac{1}{2} \cdot \left(\ell + \int_0^\infty m(s)\, ds\right)^{-3/2} \cdot \lim_{\sigma\to\infty} [-\sigma \operatorname{Im} \hat{m}(i\sigma)/\ln \sigma].$$

Now $\hat{m}(i\sigma) = \sum_{k=0}^\infty 1/(e^k + i\sigma)$; in particular,

$$\int_0^\infty m(s)\, ds = \hat{m}(0) = e/(e-1),$$

$$-\lim_{\sigma\to\infty} \sigma \operatorname{Im} \hat{m}(i\sigma)/\ln \sigma = \lim_{\sigma\to\infty} \sum_{k=0}^{\infty} \sigma^2/(e^{2k}+\sigma^2) \ln \sigma$$

$$= \lim_{\sigma\to\infty} \frac{1}{\ln \sigma} \sum_{k=0}^{\infty} 1/(e^{2k}/\sigma^2 + 1)$$

$$= \lim_{\sigma\to\infty} \frac{1}{\ln \sigma} \int_0^\infty 1/(e^{2t}/\sigma^2 + 1)\, dt$$

$$= \lim_{\sigma\to\infty} \frac{1}{2\ln \sigma} \int_{1/\sigma^2}^\infty 1/(z+1)z\, dz$$

$$= \lim_{\sigma\to\infty} \ln(1+\sigma^2)/2\ln\sigma = 1.$$

Thus $\alpha = 1/2(\ell + e/(e-1))^{3/2}$.

We will also use Example 1 to compare our work with that of Prüss [17]. We are indebted to R. Wheeler for drawing our attention to this work. The equation examined in [17] is the abstract integrodifferential equation in a Banach space X:

$$u_t(t) = \int_0^t da(t-s)Au(s)$$

where $a(t) = a_0 + a_\infty t + \int_0^t a_1(s)\,ds$ with $a_0 \geqq 0$, $a_\infty \geqq 0$ and $a_1(t)$ completely monotone. In addition, A is the generator of a cosine family in X. In order to make a reasonable comparison, we consider the equation

$$u_t(t) = \int_0^t G(t-s)Au(s)\,ds$$

where $G(t) = a_\infty + a_1(t)$.

It is shown in [17] that the resolvent operator associated with this equation becomes C^∞ in time for any $t > 0$ if

$$\lim_{\sigma\to\infty} \frac{-\ln\sigma \operatorname{Im}\hat{G}(i\sigma)}{\sigma \operatorname{Re}\hat{G}(i\sigma)} = 0.$$

If the expression above is finite but nonzero, the resolvent operator gains smoothness as time goes on.

The abstract setting can be adapted to Example 1, if nonsmooth initial conditions but homogeneous boundary conditions are considered.

In our setting we require that $G(0)$ is finite; hence $\lim_{\sigma\to\infty} i\sigma\hat{G}(i\sigma) = G(0)$. As

$$\frac{\sigma \operatorname{Im}\hat{G}'(i\sigma)}{\ln\sigma} \sim \frac{\sigma^2 \operatorname{Re}\hat{G}(i\sigma)}{\ln\sigma} \sim G(0)\frac{\sigma^2 \operatorname{Re}\hat{G}(i\sigma)}{\ln\sigma \operatorname{Re} i\sigma\hat{G}(i\sigma)}$$

$$= -G(0)\frac{\sigma \operatorname{Re}\hat{G}(i\sigma)}{\ln\sigma \operatorname{Im}\hat{G}(i\sigma)},$$

our condition for infinite smoothing is the same as that of Prüss [17].

Example 2. Our methods work as well for a system of wave equations coupled by boundary conditions. In [3], Chen, Coleman, and West have considered small transversal vibrations in a cable consisting of two or more elastic strings, linked together by a dashpot-like damping device at their ends. To find out how a step discontinuity in stress or velocity is transmitted along the cable, we must take care of the internal viscoelastic damping of the material. Assuming a linear Boltzmann constitutive law, we obtain the following system:

$$v_t^-(t,\xi) = \rho_-^{-1}\sigma_\xi^-(t,\xi), \qquad t \geqq 0, \quad \xi \in [-l, 0],$$

$$\sigma^-(t,\xi) = \int_0^t G_-(t-s)v^-(s,\xi)\,ds,$$

$$v_t^+(t,\xi) = \rho_+^{-1}\sigma_\xi(t,\xi), \qquad t \geqq 0, \quad \xi \in [0, l],$$

$$\sigma^+(t,\xi) = \int_0^t G_+(t-s)v^+(s,\xi)\,ds,$$

with the boundary conditions

$$v^-(t,-l)=v^-(t), \qquad v^+(t,l)=v^+(t),$$
$$k[v^+(t,0)-v^-(t,0)]=\sigma^+(t,0) \qquad (k>0),$$
$$\sigma^+(t,0)=\sigma^-(t,0).$$

Here $v^\pm$ denotes (transversal) velocity in either string, $\sigma^\pm$ is stress, $\rho_\pm$ is mass density, and $G_\pm$ is the stress relaxation modulus. For simplicity we have assumed both strings to have the same length l and normalized cross-section areas to one.

To fit these equations formally in our framework, we put

$$v_1(t,\xi)=v_+(t,\xi), \qquad \sigma_1(t,\xi)=\sigma_+(t,\xi),$$
$$v_2(t,\xi)=v_-(t,-\xi), \quad \sigma_2(t,\xi)=\sigma_-(t,-\xi), \quad t\geqq 0, \quad \xi\in[0,l],$$

and obtain

$$\begin{bmatrix} v_1 \\ \sigma_1 \\ v_2 \\ \sigma_2 \end{bmatrix}_t (t,\xi) = \begin{bmatrix} 0 & \rho_+^{-1} & 0 & 0 \\ G_+(0) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\rho_-^{-1} \\ 0 & 0 & -G_-(0) & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ \sigma_1 \\ v_2 \\ \sigma_2 \end{bmatrix}_\xi (t,\xi)$$
$$+\int_0^t \begin{bmatrix} 0 & 0 & 0 & 0 \\ G_+'(t-s) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -G_-'(t-s) & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ \sigma_1 \\ v_2 \\ \sigma_2 \end{bmatrix}_\xi (s,\xi)\, ds,$$

with boundary conditions

$$\begin{bmatrix} k & -1 & -k & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ \sigma_1 \\ v_2 \\ \sigma_2 \end{bmatrix} (t,0)=0,$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ \sigma_1 \\ v_2 \\ \sigma_2 \end{bmatrix} (t,l)=0.$$

The same computations as in the previous example yield

$$\lambda_1=-\sqrt{\rho_+^{-1}G_+(0)}, \qquad \lambda_2=-\sqrt{\rho_-^{-1}G_-(0)},$$
$$\lambda_3=\sqrt{\rho_-^{-1}G_-(0)}, \qquad \lambda_4=\sqrt{\rho_+^{-1}G_+(0)},$$

$$e_1=\begin{bmatrix} 1 \\ -\sqrt{\rho_+G_+(0)} \\ 0 \\ 0 \end{bmatrix}, \qquad e_2=\begin{bmatrix} 0 \\ 0 \\ 1 \\ \sqrt{\rho_-G_-(0)} \end{bmatrix},$$

$$e_3=\begin{bmatrix} 0 \\ 0 \\ 1 \\ -\sqrt{\rho_-G_-(0)} \end{bmatrix}, \qquad e_4=\begin{bmatrix} 1 \\ \sqrt{\rho_+G_+(0)} \\ 0 \\ 0 \end{bmatrix}.$$

Since

$$\begin{bmatrix} k & -1 & -k & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} [e_1, e_2] = \begin{bmatrix} k+\sqrt{\rho_+ G_+(0)} & -k \\ -\sqrt{\rho_+ G_+(0)} & -\sqrt{\rho_- G_-(0)} \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} [e_3, e_4] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

are both invertible, the boundary conditions fit in Hypothesis 2.1(b).

The further analysis depends only on the eigenvalues of

$$a + \hat{d}(\tau) = \begin{bmatrix} 0 & \rho_+^{-1} & 0 & 0 \\ \tau \hat{G}_+(\tau) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\rho_-^{-1} \\ 0 & 0 & -\tau \hat{G}_-(\tau) & 0 \end{bmatrix},$$

which is of block-diagonal form, each block corresponding to a viscoelastically damped wave equation.

To each of the blocks the results of the previous example apply. We see again that singularities in the boundary data $v^{\pm}$ are smoothed to C^{∞} in the interior of the interval, if

$$\lim_{\sigma \to \infty} \frac{\sigma \operatorname{Im} \hat{G}'_{\pm}(i\sigma)}{\ln \sigma} = \infty$$

for convex decreasing positive $-G'_{\pm}$, this means again that G' has a singularity stronger than logarithmic at 0.

If $G'(0)$ is finite, no smoothing occurs and singularities travel from end to end. It is only this case where the damping device linking the strings has influence on the propagation of singularities. Reflection and transmission occur at the connection point. The methods of [6, Ex. 3.2] show that reflection and transmission at the link are the same as if the strings were purely elastic, while the viscoelastic nature of the material leads to exponential damping of the stepsize, as waves propagate along the strings.

Example 3. Once the behavior of a single string is known, the results of the example above are in fact obvious from physical intuition. It is more interesting, however, that also a Timoshenko beam equation decouples in two wave equations, if high frequency behavior such as propagation of singularities is considered.

Consider a viscoelastic beam pinned on both ends, such that one end can be forced to move in a transversal direction at a given velocity. (Other boundary conditions may be treated similarly.) We investigate how a discontinuity in the velocity forced on the end of an initially quiescent beam is propagated along the beam. It is known that for wave phenomena with high frequencies Timoshenko's beam equation is preferable to the Euler-Bernoulli equation [21], [22]. From [21, (11.9)-(11.11)] we take

$$\rho A y_{tt} = [\kappa A G^*_{\omega}(y_{\xi} - \psi)]_{\xi},$$

$$\rho I \psi_{tt} = [E^*_{\omega} I \psi_{\xi}]_{\xi} + \kappa A G^*_{\omega}(y_{\xi} - \psi):$$

G^*_{ω} and E^*_{ω} being complex elasticity moduli. We reset the problem in the form of Boltzmann viscoelasticity, and subsequently put it in a first-order system:

$$\rho A y_{tt}(t, \xi) = \left[\kappa A G_0 (y_{\xi} - \psi)(t, \xi) + \int_0^t \kappa A G(t-s)(y_{\xi} - \psi)(s, \xi)\, ds \right]_{\xi}, \tag{4.1}$$

$$\rho I\psi_{tt}(t,\xi)=\left[E_0 I\psi_\xi(t,\xi)+\int_0^t IE(t-s)\psi_\xi(s,\xi)\,ds\right]_\xi + \kappa AG_0(y_\xi-\psi)(t,\xi)+\int_0^t \kappa AG(t-s)(y_\xi-\psi)(s,\xi)\,ds, \tag{4.2}$$

with boundary conditions

$$y_t(t,0)=v(t), \qquad y_t(t,l)=0,$$
$$\psi_\xi(t,0)=0, \qquad \psi_\xi(t,l)=0.$$

Here y is the transverse displacement of the beam, ψ is the angle of rotation for a cross-section element, when shear is neglected. G_0 and E_0 are instantaneous elasticity moduli, G and E are stress relaxation moduli. The meaning of the other constants is as in [21].

To obtain a first-order system put

$$u(t,\xi)=\begin{bmatrix} u_1(t,\xi)\\ \vdots\\ u_4(t,\xi)\end{bmatrix} \quad \text{with}$$

$$u_1=y_\xi, \quad u_2=y_t, \quad u_3=\psi_\xi, \quad u_4=\psi_t.$$

To take care of ψ in (4.2) we put $G_1(t)=G_0+\int_0^t G(s)\,ds$ and perform an integration by parts:

$$G_0\psi(t)+\int_0^t G(t-s)\psi(s)\,ds=\int_0^t G_1(t-s)\psi_t(s)\,ds.$$

(As the beam is initially quiescent, $\psi(0)=0$.) Now we can set up the first-order system:

$$u_t(t,\xi)=\begin{bmatrix} 0 & 1 & 0 & 0\\ \kappa G_0/\rho & 0 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & E_0/\rho & 0\end{bmatrix}u_\xi(t,\xi) + \begin{bmatrix} 0 & 0 & 0 & 0\\ 0 & 0 & -\kappa G_0/\rho & 0\\ 0 & 0 & 0 & 0\\ \kappa AG_0/I\rho & 0 & 0 & 0\end{bmatrix}u(t,\xi) + \int_0^t \begin{bmatrix} 0 & 0 & 0 & 0\\ \kappa G(t-s)/\rho & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & E(t-s)/\rho & 0\end{bmatrix}u_\xi(s,\xi)\,ds + \int_0^t \begin{bmatrix} 0 & 0 & 0 & 0\\ 0 & 0 & -\kappa G(t-s)/\rho & 0\\ 0 & 0 & 0 & 0\\ \kappa AG(t-s)/I\rho & 0 & 0 & -\kappa AG_1(t-s)/I\rho\end{bmatrix}u(s,\xi)\,ds,$$

$$\begin{bmatrix} 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}u(t,0)=\begin{bmatrix} v(t)\\ 0\end{bmatrix},$$

$$\begin{bmatrix} 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}u(t,l)=\begin{bmatrix} 0\\ 0\end{bmatrix}.$$

As the smoothing properties depend only on the matrices a and $d(t)$ appearing with u_ξ, we may ignore c and $h(t)$ and obtain two uncoupled viscoelastically damped wave equations exhibiting the same behavior:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}_t (t,\xi) = \begin{bmatrix} 0 & 1 \\ \kappa G_0/\rho & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}_\xi (t,\xi) + \int_0^t \begin{bmatrix} 0 & 0 \\ \kappa G(t-s)/\rho & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}_\xi (s,\xi)\, ds,$$

$$u_2(t,0) = v(t), \qquad u_2(t,l) = 0,$$

$$\begin{bmatrix} u_3 \\ u_4 \end{bmatrix}_t (t,\xi) = \begin{bmatrix} 0 & 1 \\ E_0/\rho & 0 \end{bmatrix}_\xi \begin{bmatrix} u_3 \\ u_4 \end{bmatrix} (t,\xi) + \int_0^t \begin{bmatrix} 0 & 0 \\ E(t-s)/\rho & 0 \end{bmatrix} \begin{bmatrix} u_3 \\ u_4 \end{bmatrix}_\xi (s,\xi)\, ds,$$

$$u_3(t,0) = u_3(t,l) = 0.$$

Now the computations of Example 1 show that singularities in boundary data are smoothed to C^∞ in the interior of the interval if G and E are positive, decreasing, convex with singularities stronger than logarithmic at 0.

If G and E are bounded and sufficiently smooth at $t=0$, again the methods of [6] can be applied to show that a discontinuity in boundary data for y_t and ψ_x gives rise to two waves, one carrying a discontinuity in y_t at speed $\sqrt{\kappa G_0/\rho}$, the other propagating a discontinuity in ψ_x at speed $\sqrt{E_0/\rho}$. The stepsizes are subject to exponential damping, due to the memory of the material.

Acknowledgment. W. Desch thanks Southern Illinois University for their kind hospitality.

REFERENCES

[1] J. ACHENBACH AND D. REDDY, *Note on wave propagation in linearly viscoelastic media*, Z. Angew. Math. Phys., 18 (1967), pp. 141-144.

[2] R. BIRD, R. ARMSTRONG, AND O. HASSAGER, *Dynamics of Polymeric Liquids*, John Wiley, New York, 1977.

[3] G. CHEN, M. COLEMAN, AND H. H. WEST, *Pointwise stabilization in the middle of the span for second order systems. Nonuniform and uniform exponential decay of solutions*, SIAM J. Appl. Math., to appear.

[4] R. M. CHRISTENSEN, *Theory of Viscoelasticity. An Introduction.* 2nd ed., Academic Press, New York, 1982.

[5] B. D. COLEMAN AND M. E. GURTIN, *Waves in materials with memory* II. *On the growth and decay of one-dimensional acceleration waves*, Arch. Rational Mech. Anal., 19 (1965), pp. 239-265.

[6] B. CHU, *Stress waves in isotropic linear viscoelastic materials*, J. Mécanique, 1 (1962), pp. 439-462.

[7] W. DESCH AND R. GRIMMER, *Propagation of singularities for integrodifferential equations*, J. Differential Equations, 65 (1986), pp. 411-426.

[8] ———, *Initial-boundary value problems for integrodifferential equations*, J. Integral Equations, 10 (1985), pp. 73-97.

[9] D. GRAFFI, *Mathematical models and waves in linear viscoelasticity*, in Wave Propagation in Viscoelastic Media, F. Mainardi, ed., Research Notes in Mathematics 52, Pitman, Boston, 1982.

[10] K. HANNSGEN AND R. WHEELER, *Behavior of the solution of a Volterra equation as a parameter tends to infinity*, J. Integral Equations, 7 (1984), pp. 229-237.

[11] W. HRUSA AND M. RENARDY, *On wave propagation in linear viscoelasticity*, Quart. Appl. Math., 43 (1985), pp. 237-254.

[12] J. KAZAKIA AND R. S. RIVLIN, *Run-up and spin-up in a viscoelastic fluid* I, Rheol. Acta, 20 (1981), pp. 111-127.

[13] A. NARAIN AND D. JOSEPH, *Linearized dynamics for step jumps of velocity and displacement of shearing flows of a simple fluid*, Rheol. Acta, 21 (1982), pp. 228-250.

[14] ———, *Classification of linear viscoelastic solids based on a failure criterion*, J. Elasticity, 14 (1984), pp. 19-26.

[15] A. NARAIN AND D. JOSEPH, *Linearized dynamics of shearing deformation perturbing rest in viscoelastic materials*, in Equadiff 82, H. Knobloch and K. Schmitt, eds., Lecture Notes in Mathematics 1017, Springer-Verlag, Berlin, New York, 1983.

[16] J. PRÜSS, *Positivity and regularity of hyperbolic Volterra equations in Banach spaces*, Math. Ann., 279 (1987), pp. 317-344.

[17] ———, *Regularity and integrability of resolvents of linear Volterra equations*, in Proc. Volterra Integral Equations in Banach Spaces and Applications, Trento, 1987.

[18] M. RENARDY, W. J. HRUSA, AND J. A. NOHEL, *Mathematical Problems in Viscoelasticity*, Pitman, Boston, 1987.

[19] M. RENARDY, *Some remarks on the propagation and nonpropagation of discontinuities in linearly viscoelastic fluids*, Rheol. Acta, 21 (1982), pp. 251-254.

[20a] R. RIVLIN, *Run-up and spin-up on a viscoelastic fluid* II, Rheol. Acta, 21 (1982), pp. 107-111.

[20b] ———, *Run-up and spin-up on a viscoelastic fluid* III, Rheol. Acta, 21 (1982), pp. 213-222.

[20c] ———, *Run-up and spin-up on a viscoelastic fluid* IV, Rheol. Acta, 22 (1983), pp. 275-283.

[21] J. C. SNOWDEN, *Vibration and Shock in Damped Mechanical Systems*, John Wiley, New York, 1968.

[22] S. TIMOSHENKO AND D. H. YOUNG, *Vibration Problems in Engineering*, 3rd ed., Van Nostrand, Princeton, NJ, 1955.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

010

ANALYSIS OF A CONVECTIVE REACTION-DIFFUSION EQUATION II*

H.A. LEVINE†‡, L. E. PAYNE§, P. E. SACKS†‡, AND B. STRAUGHAN¶

Abstract. We study the large time behavior of positive solutions of the semilinear parabolic equation $u_t = u_{xx} + \varepsilon(g(u))_x + f(u)$, $0 < x < L$, $\varepsilon \in \mathbf{R}$, subject to $u(0,t) = u(L,t) = 0$. The model problem in which the results apply is $g(u) = u^m$ and $f(u) = u^p$ $1 \leqq m < p$. The steady state problem is analyzed in some detail, and results about finite time blow up are proved.

Key words. Nonlinear parabolic equations, asymptotic behavior

AMS(MOS) subject classifications. 35K

Introduction. In this paper, we continue our study, begun in [1], of the longtime behavior of nonnegative solutions of

$$u_t = u_{xx} + \varepsilon(g(u))_x + f(u), \quad 0 < x < L, \quad t > 0 \tag{1.1(ε)}$$

$$u(0,t) = u(L,t) = 0, \quad t > 0 \tag{1.2}$$

$$u(x,0) = u_0(x), \quad 0 < x < L, \tag{1.3}$$

where ε is a constant ($\geqq 0$ without loss of generality) and f, g are given point functions of u. We will be primarily concerned with the power law cases $f(u) = u^p$, $g(u) = u^m$, $m, p > 1$. In the earlier paper [1], we analyzed the case $1 < p \leqq m$ and proved some general results for stationary solutions for any $p, m > 1$. We will repeat the statements of these results in the following sections. Also, in [1], we discussed some of the recent literature concerning (1.1)–(1.3). We will not repeat that discussion here.

We might point out, however, where equations of form (1.1) occur in physical situations or where (1.1) is a simplified model for a physical process. Cox and Mortell [2] show how a variety of modified Burgers' equations may be obtained by studying the equations of gas dynamics in a tube, under various boundary conditions. Horgan and Olmstead [5] investigate a system which arises from Burgers' original work and their paper contains many other relevant references. Another area is non-Boussinesq convection, where nonlinear density dependence on temperature leads to higher-order temperature effects in the momentum equation, see e.g., Veronis [12], Payne and Straughan [11].

The plan of the paper is as follows. In §2, we obtain some general propositions concerning positive stationary solutions when $p > m$. These propositions are preliminary to the main results concerning stationary solutions we obtain in §3. In that section we analyze the cases $1 < m < p < 2m - 1$ (Theorem 3.1), $p = 2m - 1$ (Theorem 3.2) and finally, $1 < 2m - 1 < p$ (Theorem 3.3). In §4, we analyze the longtime

†Department of Mathematics, Iowa State University, Ames, Iowa 50010.

‡The work of this author was supported by Air Force Office of Scientific Research grant #AFOSR 84-0252.

§The work of this author was supported in part by Sciences Engineering Research Council visiting fellowship at the University of Glasgow, Glasgow, Scotland.

¶The work of this author was partly carried out while he was a visitor at the Mathematical Sciences Institute of Cornell University, Ithaca, New York 14853.

behavior of solutions of (1.1)–(1.3). We show that when $p > m$, large data solutions cannot be global in time. Finally, in the last section, we indicate how these results may be generalized to nonlinearities that are not exact power laws.

Perhaps we should remark that when $\varepsilon = 0$, there is a close relationship between problems of the form (1.1(0)) and those of the form (1.4(0)) where for arbitrary ε,

$$u_t = u_{xx} + \varepsilon\big(g(u)\big)_x \tag{1.4(ε)}$$

$$u(0,t) = 0 \tag{1.5}$$

$$u_x(L,t) = f\big(u(L,t)\big). \tag{1.6}$$

However, when $\varepsilon \neq 0$, the sign of ε becomes crucial for the analysis of (1.4(ε)) and the correspondence between the two problems is not so close. See [7] for a discussion of (1.4(ε)).

2. Preliminary results about steady state solutions. In this section and the next, we state and prove results concerning the multiplicity of positive solutions of the steady state problem

$$\begin{aligned} &u_{xx} + \varepsilon(u^m)_x + u^p = 0, \quad 0 < x < L \\ &u(0) = u(L) = 0, \\ &u > 0, \qquad\qquad\qquad 0 < x < L \end{aligned} \tag{2.1(ε)}$$

The case $1 < p \leqq m$ has been analyzed in [1], thus we are interested here in the case $p > m \geqq 1$.

The nature of the solution set may of course depend on the parameter ε. Since a solution of (2.1(ε)) is uniquely identified by its L^∞ norm (by uniqueness of solution for the initial value problem for an ordinary differential equation) we may conveniently represent the solution of (2.1(ε)), for fixed m and p as a set of points in the $(\varepsilon, \|u\|_{L^\infty})$ plane. Also, since the change of variable $x \to L - x$ takes ε to $-\varepsilon$ we may restrict attention to $\varepsilon \geqq 0$ only. Let

$$\Gamma = \{(\varepsilon, M) : \varepsilon \geqq 0,\ M > 0 \text{ and } (2.1(\varepsilon)) \text{ has a solution } u \text{ with } \|u\|_{L^\infty} = M\}$$

be the set of points corresponding to solutions of (2.1(ε)) for some fixed values of m and p.

We begin by recalling some results from [1].

PROPOSITION 2.1. *Let $p > 1$, $m \geqq 1$.*

(i) [1, Thm. 2.2] *There is a unique positive solution of* (2.1(0)). *Denote by β_0 its L^∞ norm.*

(ii) [1, Prop. 2.6] *There exists $M_0 > 0$ such that if u solves* (2.1(ε)) *for some $\varepsilon \geqq 0$ then $\|u\|_{L^\infty} \geqq M_0$.*

(iii) [1, Prop. 2.9] *Γ is locally a simple curve.*

(iv) [1, Cor. 2.5] *In some neighborhood of $\varepsilon = 0$, $M = \beta_0$, $\Gamma = \{(\varepsilon, \beta(\varepsilon)\}$ for a continuous function β, with $\beta(0) = \beta_0$.*

(v) [1, Prop. 2.7] *The component of Γ containing $(0, \beta_0)$ is not bounded.*

(vi) [1, Lemma 2.1] *If u_1, u_2 are solutions of* (2.1(ε)), *$u_1 \not\equiv u_2$ then $u_1 < u_2$ or $u_2 < u_1$.*

Thus there is a curve of solutions emanating from the known $\varepsilon = 0$ solution. This curve may be continued indefinitely and along it either $\varepsilon \to \infty$ or $\|u\|_{L^\infty} \to \infty$ or both. We will see that for $m < p \leqq 2m - 1$ we must have $\|u\|_{L^\infty} \to \infty$ along the branch, while for $p > 2m - 1$ we must have $\varepsilon \to \infty$.

In the rest of this section we prove some results which are valid whenever $p > m \geqq 1$. We begin with some lemmas.

LEMMA 2.2. *Let u be a solution of (2.1(ε)) with $p > m - 1$. Then there exists $0 < x_0 < x_1 < L$ such that*

$$u_x \geqq 0, \quad u_{xx} \leqq 0 \quad \text{on } (0, x_0) \tag{2.2}$$

$$u_x \leqq 0, \quad u_{xx} \leqq 0 \quad \text{on } (x_0, x_1) \tag{2.3}$$

$$u_x \leqq 0, \quad u_{xx} \geqq 0 \quad \text{on } (x_1, L) \tag{2.4}$$

Proof. Clearly u_x can change sign exactly once (at some point x_0 say) and $u_{xx} \leqq 0$ on $(0, x^*)$ some $x^* > x_0$. Also since $u_x(L) < 0$ and $p > m - 1$ we must have $u_{xx} > 0$ near L, hence there exists $x_1 \in (x_0, L)$ such that $u_{xx} < 0$ on $(0, x_1)$ and $u_{xx}(x_1) = 0$. If we set $v = -u_{xx} = \varepsilon(u^m)_x + u^p$ then we obtain the following differential inequality for v;

$$\begin{aligned} v_x &= \varepsilon m u^{m-1} u_{xx} + \varepsilon m(m-1) u^{m-2} u_x^2 + p u^{p-1} u_x \\ &= \left((m-1)\frac{u_x}{u} - \varepsilon m u^{m-1}\right) v + (p - m + 1) u^{p-1} u_x \\ &< \left((m-1)\frac{u_x}{u} - \varepsilon m u^{m-1}\right) v \qquad x_1 \leqq x < L \end{aligned} \tag{2.5}$$

Since $v(x_1) = 0$ we conclude that $v(x) \leqq 0$ on $[x_1, L)$. □

LEMMA 2.3. *Let u be a solution of (2.1(ε)), $p > m - 1$. Define*

$$y(x) = \int_0^x u^{m-1}(s)\,ds \tag{2.6}$$

$$h(y) = u^m\big(x(y)\big) \tag{2.7}$$

Then

$$h_{yy} + \varepsilon m h_y + m h^{(p+1-m)/m} = 0 \tag{2.8}$$

$$h(0) = h(R) = 0 \qquad R = \int_0^L u^{m-1}(s)\,ds. \tag{2.9}$$

Furthermore if $\varepsilon_0 > 0$ there exists $R_0 > 0$, depending only on ε_0, p, m, and L such that if $0 \leqq \varepsilon \leqq \varepsilon_0$ then $R \geqq R_0$.

Proof. The fact that (2.8) holds is a straightforward computation. To prove the lower bound for R, we first claim that there exists $\alpha_0 > 0$, (depending on ε_0, p, m and L) such that

$$u_x(L) < -\alpha_0 \tag{2.10}$$

if u is a solution of (2.1(ε)) with $\varepsilon \leqq \varepsilon_0$. Allowing this for the moment, there are two possibilities. First if $x_1 < L/2$ (x_1 from Lemma 2.1) then $u(x) \geqq \alpha_0(L - x)$ for

$L/2 \leqq x \leqq L$ which clearly implies a lower bound for $\int_0^L u^{m-1}(s)ds$. If, on the other hand $x_1 > L/2$ then u is concave on $(0, x_1)$ hence is bounded below by the piecewise linear function $p(x)$ satisfying $p(0) = p(x_1) = 0$, $p(x_0) = \|u\|_{L^\infty}$. Since $\|u\|_{L^\infty}$ is bounded below by Proposition 2.1(ii) we again obtain a lower bound for R.

Finally, to prove (2.10) we see from the equation that

$$(u_x + \varepsilon u^m)_x \leqq 0 \qquad 0 < x < L \tag{2.11}$$

so that

$$u_x \geqq -\varepsilon u^m + u_x(L). \tag{2.12}$$

From this differential inequality we see that

$$\int_0^{\|u\|_{L^\infty}} \frac{du}{\varepsilon u^m - u_x(L)} \leqq L$$

so that, since $m \geqq 1$, $\|u\|_{L^\infty}$ tends to zero as $u_x(L)$ tends to zero. Hence, by Proposition 2.1(ii), $u_x(L)$ must be bounded away from zero. □

Let us denote $\psi(x) = \frac{\pi}{2L} \sin((\pi/L)\, x)$, i.e., ψ is the first eigenfunction of $-d^2/dx^2$ on $(0, L)$ normalized in $L^1(0, L)$.

LEMMA 2.4. *Let $p > m$, $\varepsilon_0 > 0$. There exists C_1 depending only on ε_0, p, m, and L such that if u is a solution of (2.1(ε)) with $\varepsilon \leqq \varepsilon_0$ then*

$$\int_0^L u\psi^{p/(p-m)} dx \leqq C_1 \tag{2.13}$$

$$\int_0^L u^p \psi^{p/(p-m)} dx \leqq C_1 \tag{2.14}$$

Proof. Let $n = \frac{p}{p-m}$, $\lambda = \left(\frac{\pi}{L}\right)^2$, multiply the equation by ψ^n and integrate from 0 to L. Simple integration by parts yields

$$\int_0^L u^p \psi^n dx + \int_0^L n(n-1) u\psi^{n-2}\psi_x^2 dx = \lambda n \int_0^L u\psi^n dx + \varepsilon n \int_0^L u^m \psi^{n-1}\psi_x \, dx. \tag{2.15}$$

Using Hölder's and Young's inequalities we find

$$\int_0^L u^p \psi^n dx \leqq \lambda n \int_0^L u\psi^n dx + \varepsilon n \left(\delta \int_0^L u^p \psi^n + C(\delta) \int_0^L |\psi_x|^n \, dx \right). \tag{2.16}$$

Choosing $\delta = 1/2\varepsilon n$, we get

$$\int u^p \psi^n \, dx \leqq 2\lambda n \int u\psi^n \, dx + C(\varepsilon_0, n, L). \tag{2.17}$$

Now from Jensen's inequality the left-hand side of (2.17) is bounded below by

$$\frac{\left(\int u\psi^n dx\right)^p}{\left(\int \psi^n dx\right)^{p-1}} \tag{2.18}$$

and the conclusion follows easily. □

LEMMA 2.5. *Let $p > m$, $\varepsilon_0 > 0$ and $0 < a < b < L$. There exists C_2 depending on $\varepsilon_0, p, m, L, a$, and b such that if u is a solution of (2.1(ε)) with $\varepsilon \leqq \varepsilon_0$ then*

$$\|u\|_{L^\infty(a,b)} \leqq C_2 \tag{2.19}$$

$$\|u_x\|_{L^1(a,b)} \leqq C_2. \tag{2.20}$$

Proof. We claim that there exists a constant $C = C(\varepsilon_0, p, m, L)$ and $x_2 \in [L/8, 7L/8]$ such that

$$|u_x(x_2)| + \varepsilon|u^m(x_2)| \leqq C. \tag{2.21}$$

Assuming (2.21) for the moment, we have from (2.1(ε)) the equation

$$u_x(x) + \varepsilon u^m(x) = u_x(x_2) + \varepsilon u^m(x_2) + \int_{x_2}^{x} u^p(s)\,ds. \tag{2.22}$$

By Lemma 2.4 u is bounded in $L^p_{\text{loc}}(0, L)$, uniformly in ε for $\varepsilon \leqq \varepsilon_0$. Hence from (2.21) and (2.22), we see that $u_x + \varepsilon u^m$ is bounded in $L^\infty_{\text{loc}}(0, L)$. Since $m < p$, u^m is bounded in $L^1_{\text{loc}}(0, L)$, again uniformly in ε for $\varepsilon \leqq \varepsilon_0$, and (2.20) follows. The estimate (2.19) follows immediately from (2.20) and (2.21).

It remains to prove the claim (2.21). From (2.13) it follows that $\text{meas}\{u\psi^n \geqq k\} < L/16$ for large enough $k = k(\varepsilon_0, p, m, L)$. Suppose u achieves its maximum at $x_0 < L/2$ (the argument is similar if $x_0 > L/2$). There must exist $x^* \in (L/2, 5L/8)$ such that $u(x^*)\psi^n(x^*) \leqq k$ and $u(x) \leqq u(x^*)$ for $x \geqq x^*$. By the mean value theorem there must exist $x_2 \in (x^*, 7L/8)$ such that

$$-u_x(x_2) = \frac{u(x^*) - u\left(\frac{7L}{8}\right)}{\frac{7L}{8} - x^*} \leqq \frac{4k}{\psi^n\left(\frac{5L}{8}\right) L} \tag{2.23}$$

Since $u(x_2) \leqq u(x^*) \leqq \frac{k}{\psi^n(5L/8)}$ the claim is proved. □

We now prove that for $p > m$ there is exactly one solution of (2.1(ε)) for ε sufficiently small. Let us denote by $\underline{u}_\varepsilon$ the known solution given by Proposition 2.1 for ε near 0.

PROPOSITION 2.6. *Let $p > m$. There exists $\varepsilon_1 > 0$ such that (2.1(ε)) has exactly one solution for $0 < \varepsilon < \varepsilon_1$.*

Proof. Suppose the conclusion is false. Then there exists $\varepsilon_j \to 0$ and u_j, a solution of (2.1ε_j), with $u_j \neq \underline{u}_{\varepsilon_j}$. Let $\zeta \in C_0^\infty(0, L)$; we have

$$\int_0^L u_j\zeta_{xx} - \varepsilon_j u_j^m\zeta_x + u_j^p\zeta\,dx = 0 \tag{2.24}$$

By Proposition 2.5 $\{u_j\}$ is precompact in $L^2_{\text{loc}}(0, L)$. Hence (passing to a subsequence) there exists $u^* \in L^2_{\text{loc}}(0, L)$ such that

$$\int_0^L [u^*\zeta_{xx} + (u^*)^p\zeta]\,dx = 0. \tag{2.25}$$

Since ζ is arbitrary we conclude that $(u^*)_{xx} \leqq 0$ weakly, and hence strongly on $(0, L)$.

Now let x_{1j} be the inflection point for u_j. We claim that $x_{1j} \to L$ as $j \to \infty$. Otherwise (for a further subsequence) there exists $L_1 < L$ such that u_j is convex on (L_1, L). Then we must have $(u^*)_{xx} \geqq 0$ on (L_1, L), a contradiction.

Hence for large enough j, u_j is concave on $(0, 3L/4)$.

We now consider two cases. If there is a subsequence $j_k \to \infty$ such that $\|u_{j_k}\|_{L^\infty} \to \infty$, then using the fact that u_{j_k} is concave on $(0, 3L/4)$ it follows that we must have $\int_0^L u_{j_k}\psi^n\,dx \to \infty$ which contradicts Lemma 2.4. If, on the other hand, $\|u_j\|_{L^\infty}$ is bounded independently of j, then one easily checks that u^* must be a solution of

(2.1(0)) or else $u^* \equiv 0$. Now $u^* \equiv 0$ is impossible because of Proposition 2.1(ii). Thus u^* must be the unique solution of (2.1(0)). This contradicts Proposition 2.1(iv) since $\underline{u}_{\varepsilon_j} \neq u_j$. □

3. Description of the set of steady state solutions In this section we complete our discussion of the steady state problem (2.1(ε)) by proving results about the multiplicity of solutions. As the exponents m and p are varied, there are at least four distinct cases which arise. Solution diagrams for these cases may be found in [1, §5].

THEOREM 3.1. *Let $1 < m < p < 2m - 1$. Then there exist $\varepsilon_0, \varepsilon_1$ such that $0 < \varepsilon_1 < \varepsilon_0 < \infty$ and such that*

(i) (2.1(ε)) *has exactly one solution for $0 \leqq \varepsilon < \varepsilon_1$;*
(ii) (2.1(ε)) *has no solution for $\varepsilon > \varepsilon_0$;*
(iii) *If p is close enough to m then there exists ε such that* (2.1(ε)) *has at least two solutions, i.e. $\varepsilon_1 < \varepsilon_0$.*

THEOREM 3.2. *Let $p = 2m - 1$, $m > 1$. Then*

(i) (2.1(ε)) *has exactly one solution for $0 \leqq \varepsilon < \varepsilon_0 \equiv \sqrt{\frac{4}{m}}$;*
(ii) (2.1(ε)) *has no solution for $\varepsilon \geqq \varepsilon_0$.*

THEOREM 3.3. *Let $p > 2m - 1$. Then* (2.1(ε)) *has exactly one solution for all $\varepsilon \geqq 0$.*

As one can see, the results are not complete for $m < p < 2m - 1$. We conjecture that $\varepsilon_1 < \varepsilon_0$ for all m, p in this range.

For the readers convenience we recall the result of [1] for the case $p \leqq m$.

THEOREM 3.4. *Let $1 < p \leqq m$. Then there exists $\varepsilon_0 > 0$ such that*

(i) (2.1(ε)) *has at least two solutions for $0 < \varepsilon < \varepsilon_0$;*
(ii) (2.1(ε)) *has at least one solution for $\varepsilon = \varepsilon_0$;*
(iii) (2.1(ε)) *has no solution for $\varepsilon > \varepsilon_0$.*

We begin the proofs with the easiest case, $p = 2m - 1$. Note that this is exactly the case in which equation (2.8) is linear.

Proof of Theorem 3.2. Define y, h, and R as in Lemma 2.3. If $\varepsilon \geqq \sqrt{4/m}$ all solutions of (2.8) are nonoscillatory, hence $h(0) = h(R) = 0$ cannot be satisfied for any $R > 0$ except if $h \equiv 0$.

To show the existence of at least one solution for $\varepsilon < \sqrt{4/m}$, then by Proposition 2.1 it is enough to obtain an a priori bound for solutions of (2.1(ε)) when $\varepsilon < \sqrt{4/m}$. Thus suppose u is a solution for such an ε. We have by explicit calculation

$$h(y) = Ce^{-(\varepsilon m/2)y} \sin\left(\sqrt{m - \frac{\varepsilon^2 m^2}{4}}\, y\right) \tag{3.1}$$

for some constant C. Since $h > 0$

$$\int_0^L u^{m-1}(s)\,ds = R = \frac{\pi}{\sqrt{m - \frac{\varepsilon^2 m^2}{4}}}. \tag{3.2}$$

The constant C can be determined from

(3.3)

$$y' = u^{m-1} = h^{(m-1)/m} = C^{(m-1)/m} e^{\varepsilon((m-1)/2)y} \sin^{(m-1)/m}\left(\left(\sqrt{m - \frac{\varepsilon^2 m^2}{4}}\right) y\right);$$

i.e.,

$$C^{(m-1)/m} = \frac{1}{L}\int_0^R \frac{e^{\varepsilon\left((m-1)/2\right)y}\,dy}{\sin^{(m-1)/m}\left(\left(\sqrt{m - \frac{\varepsilon^2 m^2}{4}}\right) y\right)}. \tag{3.4}$$

Hence

$$\|u\|_{L^\infty} \leqq C^{m-1} \tag{3.5}$$

and C clearly remains bounded when ε is bounded away from $\sqrt{4/m}$.

Finally, if for some ε there exists two solutions $u_1 \not\equiv u_2$ then the preceding argument shows $\int_0^L u_1^{m-1}\,dx = \int_0^L u_2^{m-1}\,dx$. However, by Proposition 2.1(vi) $u_1 < u_2$ or $u_2 < u_1$, a contradiction. □

Remark. The uniqueness of positive solutions of (2.1(ε)) when $p = 2m - 1$ can also be shown by a scaling argument as follows. The differential equation is invariant under the transformation $u(x) \to \lambda^{1/(p-m)} u(\lambda x)$ for any $\lambda > 0$. Thus, if there existed two distinct positive solutions of (2.1(ε)) then we could find two functions u_i $i = 1, 2$ with $u_1(0) = u_2(0) = 0$, $\|u_1\|_{L^\infty} = \|u_2\|_{L^\infty}$ and both satisfying $u_{xx} + \varepsilon(u^m)_x + u^p = 0$. By uniqueness of solutions of the initial value problem for u, we may conclude that $u_1 \equiv u_2$.

For the proof of Theorem 3.3 we need the following.

LEMMA 3.4. *Let $p > 2m - 1$. Fix $\varepsilon \geqq 0$ and let h_i be positive solutions of* (2.8), (2.9) *on* $[0, R_i]$ *for* $i = 1, 2$. *If* $R_2 > R_1$ *then* $\|h_2\|_{L^\infty} < \|h_1\|_{L^\infty}$.

Proof. If $\|h_2\|_{L^\infty} = \|h_1\|_{L^\infty}$ then $h_1 \equiv h_2$ by uniqueness of solutions of the initial value problem and hence $R_1 = R_2$. Now suppose $\|h_2\|_{L^\infty} > \|h_1\|_{L^\infty}$. By consideration of the phase plane diagram for (2.8) we see that

$$h_2'(R_2) < h_1'(R_1) < 0 < h_1'(0) < h_2'(0). \tag{3.6}$$

We first observe that it is not possible that $h_2(y) \geqq h_1(y)$ for $0 \leqq y \leqq R_1$. Indeed, otherwise

$$(h_1 h_2' - h_2 h_1')' + \varepsilon(h_1 h_2' - h_2 h_1') = h_2 h_1^{(p-m+1)/m} - h_1 h_2^{(p-m+1)/m} < 0 \tag{3.7}$$

since $p > 2m - 1$, so that

$$h_1 h_2' - h_2 h_1' \leqq \big(h_1(0)h_2'(0) - h_2(0)h_1'(0)\big)e^{-\varepsilon y} = 0. \tag{3.8}$$

and in particular

$$-h_2(R_1)h_1'(R_1) \leqq 0,$$

which is a contradiction.

Thus there must exist $y_1 \in (0, R_1)$ such that $h_2 > h_1$ on $(0, y_1)$ and $h_2(y_1) = h_1(y_1)$. There are now two cases. First suppose $h_2'(y_1) \geqq 0$.

Since (3.8) holds on $(0, y_1]$ it follows that $h_1'(y_1) > h_2'(y_1)$. Denote by h_i^{-1} the branches of the inverse function with values in $[0, y_1]$. We see that $h_1'\big(h_1^{-1}(x)\big) - h_2'\big(h_2^{-1}(z)\big)$ must have a root in $\big(0, h_1(y_2)\big)$, i.e., there exist points y_2, y_2^* such that $h_1(y_2) = h_2(y_2^*)$ and $h_1'(y_2) = h_2'(y_2^*)$. Thus by uniqueness for the initial value problem, $h_1(y) = h_2(y + y_2^* - y_2)$, and since $h_1(0) = h_2(0) = 0$ we must have $h_1 \equiv h_2$, a contradiction.

Finally suppose $h_2'(y_1) < 0$. Since $h_2(R_1) > h_1(R_1) = 0$ there must exist $y_2 \in (y_1, R_1)$ such that $h_2(y) > h_1(y)$ on (y_2, R_1) and $h_2(y_2) = h_1(y_2)$. Also we must have $h_1'(y_2) < 0$ and $h_1'(y_2) < h_2'(y_2) < 0$. Using (3.6) we again have a contradiction as in the previous case. □

Proof of Theorem 3.3. First we prove uniqueness. Suppose u_1 and u_2 are both solutions of (2.1(ε)); by Proposition 2.1(vi) we may assume $0 < u_1 < u_2$. Let h_i, R_i $i = 1, 2$ be as in Lemma 2.3, $i = 1, 2$, so that $R_2 > R_1$. Then, from Lemma 3.4, it follows that $\|u_2\|_{L^\infty} < \|u_1\|_{L^\infty}$ a contradiction.

We next show the existence of a solution for all $\varepsilon \geqq 0$. By Proposition 2.1(v) it is enough to show that if $\varepsilon_0 > 0$ there exists $C = C(\varepsilon_0, p, m, L)$ such that if u is a solution of (2.1(ε)) for $\varepsilon \leqq \varepsilon_0$, then $\|u\|_{L^\infty} \leqq C$. Using Lemmas 3.4 and 2.3 it is enough to prove the same statement for solutions of (2.8), (2.9) for a fixed $R = R_0 > 0$.

Let $\psi_\varepsilon(y)$, $\lambda_\varepsilon > 0$ satisfy

$$\psi_\varepsilon'' - \varepsilon m \psi_\varepsilon' = -\lambda_\varepsilon \psi_\varepsilon \tag{3.9}$$

$$\psi_\varepsilon(0) = \psi_\varepsilon(R_0) = 0 \tag{3.10}$$

$$\int_0^{R_0} \psi_\varepsilon(y)\, dy = 1. \tag{3.11}$$

Multiplying (2.8) by ψ_ε and integrating gives

$$\int_0^{R_0} \psi_\varepsilon(y) h^{(p+1-m)/m}(y)\, dy = \frac{\lambda_\varepsilon}{m} \int \psi_\varepsilon(y) h(y)\, dy. \tag{3.12}$$

Hence using Jensen's inequality,

$$\int_0^{R_0} \psi_\varepsilon(y) h^{(p+1-m)/m}(y)\, dy \leqq C(\varepsilon_0, p, m, L). \tag{3.13}$$

Now let $G_\varepsilon(y, \xi)$ be the Green's function for

$$\varsigma'' + \varepsilon m \varsigma' = f \qquad 0 < y < R_0 \tag{3.14}$$

$$\varsigma(0) = \varsigma(R_0) = 0 \tag{3.15}$$

so that

$$h(y) = m \int_0^{R_0} G_\varepsilon(y, \xi) h^{(p+1-m)/m}(\xi)\, d\xi. \tag{3.16}$$

It is a straightforward calculation to check that

$$\left| \frac{G_\varepsilon(y, \xi)}{\psi_\varepsilon(\xi)} \right| \leqq C(\varepsilon_0, m, R_0) \qquad \varepsilon \leqq \varepsilon_0. \tag{3.17}$$

Hence, from (3.13) and (3.17), one obtains the required uniform estimate for h. □

For the proof of Theorem 3.1 we begin with the nonexistence result.

PROPOSITION 3.5. *Let $1 < m < p < 2m - 1$. Then there exists $\varepsilon_0 > 0$ such that (2.1(ε)) has no solution for $\varepsilon \geqq \varepsilon_0$.*

Proof. First we make the following claim. There exists $\alpha_0 > 0$, $\varepsilon^* > 0$ such that if u is a solution of (2.1(ε)) for $\varepsilon > \varepsilon^*$ then there exists $x_2 > x_1$ (x_1 from Lemma 2.2) such that

$$u(x_2) \geqq \alpha_0 \tag{3.18}$$

$$mu_x(x_2) \leqq -\frac{2}{\varepsilon}u^{p-m+1}(x_2) \tag{3.19}$$

To see this define

$$\tilde{x}_2 = \sup\left\{x : mu_x(\hat{x}) + \frac{2}{\varepsilon}u^{p-m+1}(\hat{x}) \geqq 0 \quad \hat{x} < x\right\}. \tag{3.20}$$

Since

$$mu_x(x) + \frac{1}{\varepsilon}u^{p-m+1}(x) \geqq 0 \qquad x \leqq x_1$$

and

$$mu_x(L) + \frac{1}{\varepsilon}u^{p-m+1}(L) < 0,$$

$\tilde{x}_2$ is well defined, and $\tilde{x}_2 \in (x_1, L)$. Now recalling from Proposition 2.1(ii) that $u(x_0) \geqq M_0 > 0$ and integrating the differential inequality $mu_x + (2/\varepsilon)\,u^{p-m+1}(x) \geqq 0$ on $[x_0, \tilde{x}_2]$, we find that $u(\tilde{x}_2) \geqq M_0/2$ for say $\varepsilon > \varepsilon^*$. Choosing $\alpha_0 = M_0/4$ we have that the conditions (3.18) and (3.19) are satisfied by some $x_2 > \tilde{x}_2$.

Next define the function $h(y)$ as in Lemma 2.3 and let $\varsigma(y)$ satisfy

$$\varsigma'' + \varepsilon m\varsigma' + mh(y_2)^{(p+1-2m)/m}\varsigma = 0 \tag{3.21}$$

$$\varsigma(y_2) = h(y_2) \tag{3.22}$$

$$\varsigma'(y_2) = h'(y_2), \tag{3.23}$$

where $y_2 = y(x_2)$. The inequalities (3.18), (3.19) in terms of h are

$$h(y_2) \geqq \alpha_0^m \tag{3.24}$$

$$h'(y_2) < -\frac{2}{\varepsilon}h^{\frac{p-m+1}{m}}(y_2) \tag{3.25}$$

Let $\varepsilon^2 > \max(\varepsilon^{*2}, (4/m)\alpha_0^{p+1-2m})$. We claim that $\varsigma'(y) \leqq 0$ for $y \leqq y_2$. Accepting this for the moment, we have (since $p + 1 - 2m < 0$)

$$(h\varsigma' - h'\varsigma)' + \varepsilon m(h\varsigma' - h'\varsigma) = -m\varsigma h\left(h(y_2)^{(p+1-2m)/m} - h^{(p+1-2m)/m}\right). \tag{3.26}$$

There is an interval $[y_3, y_2]$ on which $h(y) \geqq h(y_2) = h(y_3)$ and so

$$[e^{\varepsilon m y}(h\varsigma' - h'\varsigma)]' \leqq 0 \qquad \text{on } [y_3, y_2] \tag{3.27}$$

and in particular

$$h(y)\varsigma'(y) - h'(y)\varsigma(y) \geqq 0 \qquad y_3 \leqq y \leqq y_2. \tag{3.28}$$

But the interval $[y_3, y_2]$ contains the point $y(x_0)$, where the maximum of h is achieved, and evaluating (3.28) at this point we obtain $\varsigma'(y(x_0)) > 0$, a contradiction.

It remains to verify that $\varsigma'(y) \leqq 0$ for $y \leqq y_2$. This may be checked by writing out explicitly the solution $\varsigma(y)$, and using (3.21)–(3.25). The computation can be made a little less painful by arguing in the following way. First, the characteristic roots for the ς equation are real, hence ς can have no more than one critical point. Since $\varsigma'(y_2) < 0$, if we check that $\varsigma'(y) < 0$ for large negative y we will be done. Now as $y \to -\infty$ it is easy to check that $\varsigma(y) \sim Ce^{ry}$ where $r = -(A + \sqrt{A^2 - 4B})/2$, $C = (rh(y_2) - h'(y_2))/\sqrt{A^2 - 4B}$ and $A = \varepsilon m$, $B = mh(y_2)^{(p+1-2m)/m}$. Thus to conclude we need $h'(y_2) < rh(y_2)$, and because of (3.25) it is enough that $2/\varepsilon\, h^{(p+1-2m)/m}(y_2) \geqq A/2[1 - \sqrt{1 - 4B/A^2}]$. But the left side is $2B/A$, and since $4B/A^2 < 1$ the conclusion follows from the obvious inequality $1 - \sqrt{1 - \xi} < \xi$ for $0 < \xi < 1$. □

Proof of Theorem 3.1. We have already proved (i) in Proposition 2.6 and (ii) in Proposition 3.5. For the proof of (iii) we will need to recall some results from [1]. In that paper solutions of (2.1(ε)) were found by looking for zeros of a function $H(\alpha, \varepsilon) = v(L)$, where $v(x)$ satisfies

$$v_{xx} + \varepsilon(v^m)_x + v^p = 0 \qquad x > 0 \tag{3.29}$$

$$v(0) = 0 \tag{3.30}$$

$$v_x(0) = \alpha. \tag{3.31}$$

Let us fix m and regard H as a function of p also, $H = H(\alpha, \varepsilon, p)$. For $p > 1$ H is C^1 in all variables and by Proposition 2.9 of [1] either $\partial H/\partial\alpha$ or $\partial H/\partial\varepsilon$ is nonzero at any root of H. Also by Proposition 2.4 of [1] $\partial H/\partial\alpha(\alpha_0, 0) > 0$, where α_0 is the initial slope for the unique solution of (2.1(0)). Thus it is not hard to check that there exists $\overline{\varepsilon} > 0$, $\overline{\alpha} > \alpha_0$ and $\overline{m} > m$ such that for $p \in [m, \overline{m}]$, (2.1(ε)) has a solution for all $\varepsilon \in [0, \overline{\varepsilon}]$ with initial slope $\alpha = \alpha(\varepsilon) \leqq \overline{\alpha}$.

Now for $p = m$ it was shown in Proposition 3.5 of [1] that (2.1(ε)) has at least two solutions for all $0 < \varepsilon < \varepsilon_0$, and in particular there is a maximal solution $\overline{u}_\varepsilon$ of (2.1(ε)) with $\overline{u}_{\varepsilon x}(0) \to \infty$ as $\varepsilon \to 0$. Choosing $\varepsilon_2 \in (0, \overline{\varepsilon})$ such that $\overline{u}_{\varepsilon_2 x}(0) > \overline{\alpha}$ yields a curve of roots of H in the form $\alpha = \alpha(\varepsilon, p)$ or $\varepsilon = \varepsilon(\alpha, p)$ containing the point $(u_{\varepsilon_2 x}(0), \varepsilon_2, m)$ and hence for p close enough to m there must be a second solution of (2.1(ε)) for some ε near ε_2. □

4. The time dependent problem. In this section we discuss the large time behavior for the time dependent problem

$$\begin{aligned} (4.1(\varepsilon)) \qquad u_t &= u_{xx} + \varepsilon(u^m)_x + u^p \qquad 0 < x < L \qquad t > 0 \\ u(0,t) &= u(L,t) = 0 \qquad\qquad\qquad\quad t > 0 \\ u(x,0) &= u_0(x) \geqq 0 \qquad\quad 0 < x < L \end{aligned}$$

It is known (e.g., [1], [8]) that for any $u_0 \in L^\infty(0, L)$ there is a solution $u(x,t)$ of (4.1(ε)) on some interval $0 \leqq t \leqq T_{\max}$, for some $T_{\max} \in (0, \infty]$, and if $T_{\max} < \infty$ then $\lim_{t \to T_{\max}^-} \|u(\cdot, t)\|_{L^\infty(0,L)} = \infty$.

Our first result gives sufficient conditions that $T_{\max}$ be finite, i.e. that the solution of (4.1(ε)) blow up in finite time. Recall we have defined $\psi(x) = (\pi/2L)\, sin\, ((\pi/L)x)$ and $n = p/(p - m)$.

PROPOSITION 4.1. *Let $p > m$. Then there exists $C_0 = C_0(\varepsilon, p, m, L) < \infty$ such that if*

$$\int_0^L u_0(x)\psi^n(x)\,dx > C_0 \tag{4.2}$$

then the solution of (4.1(ε)) blows up in finite time.

Remark. The proof is a variation of the eigenfunction method (e.g., Kaplan [6], Levine [14] and Payne [10]). The nature of the blow-up set has been studied by Friedman and Lacey [4]. Recall from [1] that if $\varepsilon \neq 0$ and $p \leqq m$ then $T_{\max} = \infty$ for any $u_0 \in L^\infty(0, L)$. See Chipot and Weissler [3] for another recent result about blow-up for an equation with nonlinear gradient dependence; previous work on blow-up has generally only dealt with the reaction-diffusion case.

Proof. Set

$$F(t) = \int_0^L u(x,t)\psi^n(x)\,dx. \tag{4.3}$$

A short calculation shows that

$$F'(t) \geqq -(\pi^2/L^2)\,n\int_0^L \psi^n(x)u(x,t)\,dx - \varepsilon n\int_0^L u^m(x,t)\psi^{n-1}(x)\psi_x(x)\,dx + \int_0^L u^p(x,t)\psi^n(x)\,dx. \tag{4.4}$$

Using Jensen's inequality we see that

$$\varepsilon n\int_0^L u^m(x,t)\psi^{n-1}(x)\psi_x(x)\,dx \leqq A_1\left(\int_0^L u^p(x,t)\psi^n(x)\,dx\right)^{m/p} \tag{4.5}$$

$$\frac{\pi^2}{L^2}n\int_0^L u(x,t)\psi^n(x)\,dx \leqq A_2\left(\int_0^L u^p(x,t)\psi^n(x)\,dx\right)^{1/p} \tag{4.6}$$

with

$$A_1 = \varepsilon n\|\psi_x\|_{L^\infty(0,L)}L^{1-m/p} \tag{4.7}$$

$$A_2 = \frac{\pi^2}{L^2}n\left(\int_0^L \psi^n(x)\,dx\right)^{(p-1)/p}. \tag{4.8}$$

Therefore

$$F'(t) \geqq Q\left(\int_0^L u^p(x,t)\psi^n(x)\,dx\right) \tag{4.9}$$

with

$$Q(s) = s - A_1 s^{m/p} - A_2 s^{1/p}. \tag{4.10}$$

If s_0 is the largest positive root of Q, then $Q(s), Q'(s) > 0$ for $s > s_0$. If

$$F(0) > C_0 \equiv A_2 s_0^{1/p}, \tag{4.11}$$

then one sees easily that $\int_0^L u^p(x,t)\psi^n(x)\,dx > s_0$ for all t and so

$$F'(t) \geqq Q\left(\frac{1}{A_2}F^p(t)\right). \tag{4.12}$$

Since $p > 1$, $F(t)$ must blow up in finite time. □

We next recall some results from [1].

PROPOSITION 4.2. *Let $m \geqq 1$, $p > 1$, $\varepsilon \in \mathbf{R}$*

(i) *$u \equiv 0$ is an asymptotically stable steady state solution of (4.1(ε)).*
(ii) *If (4.1(ε)) has any positive steady state then it has a minimal one, $\underline{u}_\varepsilon$, which is unstable from below.*
(iii) *Let $E = \{\varepsilon : \underline{u}_\varepsilon$ is stable from above$\}$. Then E is nowhere dense and there exists $\varepsilon_2 > 0$ such that $E \cap (-\varepsilon_2, \varepsilon_2) = \varnothing$.*

Remark. In general the possibility that a steady state is stable from above and unstable from below (i.e., is not hyperbolic) cannot be ruled out, such behavior being necessary at the "turning points" of the solution curve Γ (defined at the beginning of §2), which may occur for $1 < p \leqq m$.

Another important result concerning the asymptotic behavior of solutions of (4.1(ε)) follows from general theorems due to Zelenyak [13].

PROPOSITION 4.3. *Let $u(x,t)$ be a solution of (4.1(ε)) such that $\|u(\cdot,t)\|_{L^\infty(0,L)}$ is uniformly bounded for $t > 0$. Then there exists a nonnegative steady state u^* of (4.1(ε)) such that $u(x,t) \to u^*(x)$ uniformly as $t \to \infty$.*

Remark. We may conclude, for example, from Propositions 4.2(ii), 4.3 and the maximum principle that if $u_0(x) \leqq \underline{u}_\varepsilon(x)$, $u_0(x) \not\equiv \underline{u}_\varepsilon(x)$ then the solution of (4.1(ε)) tends to 0 as $t \to \infty$.

To conclude we summarize all that we can say about the asymptotic behavior of solutions of (4.1(ε)) under various conditions on m, p, and ε. In what follows, the numbers ε_0, ε_1, and ε_2 have the meaning assigned to them in Theorems 3.1, 3.2, 3.3, and Proposition 4.2.

Case 1. $m < p < 2m - 1$ and $\varepsilon > \varepsilon_0$ or $p = 2m - 1$ and $\varepsilon \geqq \varepsilon_0$. In this case the only nonnegative steady state is $u \equiv 0$, which is asymptotically stable. If $u(x,t)$ is the solution of (4.1(ε)) then either it converges to 0 as $t \to \infty$ or it tends to infinity in finite or infinite time. If u_0 is small enough (in $L^\infty(0,L)$ say) then u must tend to zero, while if u_0 is large enough so that 4.2 holds, then u must go to ∞ in finite time. We mention that Matano has shown ([9]), for the case of a reaction-diffusion equation in one space dimension, that blow up cannot occur if the maximal nonnegative solution is stable from above. The present example shows that Matano's result is false in general if the equation has some first derivative terms.

Case 2. $m < p < 2m - 1$ and $\varepsilon < \varepsilon_1$ or $p = 2m - 1$ and $\varepsilon < \varepsilon_0$ or $p > 2m - 1$. In this case there is exactly one positive steady state $\underline{u}_\varepsilon$ which is unstable from below. For any u_0 the solution must either tend to 0 as $t \to \infty$ or tend to $\underline{u}_\varepsilon$ as $t \to \infty$ or tend to infinity in finite or infinite time. If $u_0 \leqq \underline{u}_\varepsilon$, $u_0 \not\equiv \underline{u}_\varepsilon$ or if u_0 is sufficiently small in $L^\infty(0,L)$ then $u \to 0$ as $t \to \infty$. If u_0 is large enough so that (4.2) holds then u tends to infinity in finite time.

Case 3. Same as Case 2 and also $\varepsilon \notin E$ (in particular $0 < \varepsilon < \varepsilon_2$). In this case we also have $\underline{u}_\varepsilon$ unstable from above so we can assert in addition that if $u_0 \geqq \underline{u}_\varepsilon$, $u_0 \not\equiv \underline{u}_\varepsilon$, then u tends to infinity in finite or infinite time.

Case 4. $m < p < 2m-1$, $\varepsilon_1 < \varepsilon < \varepsilon_0$. In this case there are one or more positive steady states. For any u_0, the solution u must tend to a nonnegative steady state or tend to infinity in finite or infinite time. Again if $u_0 \leqq \underline{u}_\varepsilon$, $u_0 \not\equiv \underline{u}_\varepsilon$ or if u_0 is small enough in $L^\infty(0,L)$ then $u \to 0$ as $t \to \infty$ while if 4.2 holds then u tends to infinity in finite time. If $\varepsilon \notin E$ then there must exist another steady state solution $\overline{u}_\varepsilon > \underline{u}_\varepsilon$ which is stable from below.

5. More general nonlinearities. In this section we describe briefly how the results of §§2-4 can be generalized to the case of nonpower-law nonlinearities. We will write the steady state problem as

$$(5.1\varepsilon) \qquad \begin{aligned} u_{xx} + \varepsilon\big(g(u)\big)_x + f(u) &= 0 \qquad 0 < x < L \\ u(0) = u(L) &= 0 \\ u(x) &> 0 \qquad 0 < x < L. \end{aligned}$$

Let us first list all of the hypotheses that may be used

(H1) $f \in C^1\big([0,\infty)\big)$, $f(0) = 0$, $f(u)/u$ is strictly increasing on $\mathbf{R}^+$

$$0 \leqq \lim_{u\to 0} \frac{f(u)}{u} < \left(\frac{\pi}{L}\right)^2 < \lim_{u\to\infty} \frac{f(u)}{u}.$$

(H2) $g \in C^1\big([0,\infty)\big)$, $g(0) = 0$, $g'(u) > 0$ for $u > 0$.

(H3) $f(u)/g'(u)$ is nondecreasing on $\mathbf{R}^+$.

(H4) $f(u)/g(u)g'(u)$ is strictly increasing on $\mathbf{R}^+$ with $\lim_{u\to\infty} f(u)/g(u)g'(u) = \infty$.

(H5) $f(u)/g(u)g'(u)$ is strictly decreasing on $\mathbf{R}^+$ with $\lim_{u\to\infty} f(u)/g(u)g'(u) = 0$.

(H6) $f(u)/g(u)g'(u) \equiv \lambda$ (a constant).

(H7) $f(u) \geqq C_1 u^p - C_2$, $u \geqq 0$, some $C_1, C_2 \geqq 0$.

(H8) $g(u) \leqq C_3 u^m + C_4$, $u \geqq 0$, some $C_3, C_4 \geqq 0$.

THEOREM 5.1. *Let (H1), (H2) and (H3) hold*

(i) *Assume* (H5), (H7) *and* (H8) *hold with* $p > m$. *Then there exists* $\varepsilon_0, \varepsilon_1$ *such that* $0 < \varepsilon_1 \leqq \varepsilon_0 < \infty$ *and such that* (5.1(ε)) *has exactly one solution for* $0 \leqq \varepsilon < \varepsilon_1$ *and no solution for* $\varepsilon > \varepsilon_0$.

(ii) *Assume* (H6). *Then* (5.1(ε)) *has exactly one solution for* $0 \leqq \varepsilon < \sqrt{4\lambda}$ *and no solution for* $\varepsilon \geqq \sqrt{4\lambda}$.

(iii) *Assume* (H4). *Then* (5.1(ε)) *has exactly one solution for all* $\varepsilon \geqq 0$.

Let us make some remarks about which hypotheses are necessary for the various Lemmas and Propositions. Proposition 2.1 is true if the conditions $p > 1$ and $m \geqq 1$ are replaced by (H1) and (H2), except possibly for part (iii) which is proved in [1] only for the case that (H1) holds and $g(u) = u^m$, $m \geqq 1$. However, the fact that Γ is locally a simple curve is not explicitly used in what follows. The conclusion of Lemma 2.2 is still true if we assume (H1), (H2), and (H3). In Lemma 2.3 we set

$$(5.2) \qquad y(x) = \int_0^x g'\big(u(s)\big)ds$$

$$(5.3) \qquad h(y) = g\big(u(y)\big)$$

and in place of (2.8) we have

$$(5.4) \qquad h_{yy} + \varepsilon h_y + F(h) = 0$$

with

$$F(h) = f(g^{-1}(h))/g'(g^{-1}(h)). \tag{5.5}$$

The lower bound for $R = \int_0^L g'(u(s))ds$ in Lemma 2.3 remains valid if we assume (H1), (H2), and (H3). Lemmas 2.4 and 2.5 are true if we assume (H7) and (H8) with $p > m$; in (2.14), we replace u^p by $f(u)$. The conclusion of Proposition 2.6 still holds if we assume (H1), (H2), (H3), (H7), and (H8) with $p > m$. In §3, Lemma 3.4 will be true under hypotheses (H1), (H2), and (H4), and for Propositon 3.5 we assume (H1), (H2), (H3), and (H5).

Concerning the time dependent problem

$$\begin{aligned} (5.6(\varepsilon)) \qquad & u_t = u_{xx} + \varepsilon(g(u))_x + f(u) && 0 < x < L \qquad t > 0 \\ & u(0,t) = u(L,t) = 0 \\ & u(x,0) = u_0(x) \geqq 0 && 0 < x < L \end{aligned}$$

an adequate existence, uniqueness, and continuation theory will be true, provided (H1) and (H2) hold, and an analogue of the blow-up result Proposition 4.1 is proved as before if we assume also H7 and H8 with $p > m$.

In Proposition 4.2, parts (i) and (ii) are true assuming (H1) and (H2). The set E in part (iii) still cannot intersect an interval $(-\varepsilon_2, \varepsilon_2)$ but the fact that it is nowhere dense was proved in [1] assuming only that (H1) holds and $g(u) = u^m$ $m \geqq 1$. The stabilization result, Proposition 4.2, is true under conditions (H1) and (H2), although more regularity is assumed in [13]. We leave to the interested reader the formulation of results about the asymptotic behavior analogous to those stated at the end of §4.

REFERENCES

[1] T. F. CHEN, H. A. LEVINE,and P. E. SACKS, *Analysis of a convective reaction diffusion equation*, J. Nonlinear Anal. T.M.A., to appear.

[2] E. A. COX and M. P. MORTELL, *The evolution of resonant oscillations in closed tubes*, ZAMP, 34 (1983), pp. 845–866.

[3] M. CHIPOT and F. B. WEISSLER, *Some blow up results for a nonlinear parabolic equation with a gradient term*, preprint.

[4] A. FRIEDMAN and A. A. LACEY, *Blow up of solutions of semilinear parabolic equations*, J. Math. Anal. Appl., to appear.

[5] C. O HORGAN and W. E. OLMSTEAD, *Stability and uniqueness for a turbulence model of Burgers*, Quart. Appl. Math., 36 (1978), pp. 131–127.

[6] S. KAPLAN, *On the growth of solutions of quasilinear parabolic equations*, Comm. Pure Appl. Math., 16 (1963), pp. 305–330.

[7] H. A. LEVINE, *Stability and instability for solutions of Burgers' equation with a semilinear boundary condition*, SIAM J. Math. Anal., 19 (1988), pp. 312–336.

[8] O. A. LADYZENSKAYA, V. A. SOLONNIKOV, and N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, R.I., 1967.

[9] H. MATANO, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Pub. R.I.M.S., 15 (1975), pp. 401–454.

[10] L. E. PAYNE, *Improperly Posed Problems in Partial Differential Equations*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.

[11] L. E. PAYNE and B. STRAUGHAN, *Unconditional nonlinear stability in penetrative convection*, Geophys. Astrophys. Fluid Dyn., to appear.

[12] G. VERONIS, *Penetrative Convection*, Astrophys. J., 137 (1963), pp. 641–663.

[13] T. I. ZELENYAK, *Stabilization of solutions of boundary value problems for a second order parabolic equation with one space variable*, Differential Equations, 4 (1968), pp. 17–22.

[14] H. A. LEVINE, *On the nonexistence of global weak solutions of some properly and improperly posed problem of mathematical physics: The methods of unbounded Fourier coefficients*, Math. Annalen, 214 (1975), pp. 205–220.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

011

NONMONOTONE CLINES IN HOMOGENEOUS SPACE CANNOT BE STABLE*

EBBE THUE POULSEN†

Abstract. This paper deals with models for the space-time variation of the genetic composition of populations. Both the standard reaction-diffusion equation and two nonlinear integral operator models are discussed. A simple argument is presented, explaining that nonconstant stationary solutions, so-called *clines*, in a homogeneous space cannot be stable unless they are monotone.

In the integral operator case the proof uses differentiation along the orbits of the translation group and the Perron–Frobenius Theorem for integral operators with nonnegative kernel. For the reaction-diffusion equation stronger results are known, and only a heuristic argument is given.

Key words. clines, instability, equilibria, discrete-time dynamical systems, Perron-Frobenius property, nonlinear integral operators

AMS(MOS) subject classifications. primary 45M10; secondary 92A10.

1. Some models in population genetics and dynamics. Population genetics has given rise to an abundance of mathematical models distinguished by various characteristic properties: deterministic or stochastic models; continuous or discrete time; haploid or diploid species; one locus or many loci, and, in the one-locus-case: two or more alleles; one habitat, several discrete habitats, or a habitat that is a region in $\mathbf{R}^n$ for $n=1$, 2, or 3. The mathematical problems considered in this paper correspond to deterministic models for one locus with two alleles (denoted A and a) in a diploid population distributed over all of $\mathbf{R}$.

As the literature amply testifies, even the very simplest problem of a population with one locus and two alleles in just one habitat has given rise to an impressive number of different models, and for some of these the analysis of the dynamics is quite complicated. Part of the reason for this complexity is the fact that the natural state space is three-dimensional, the three natural variables being the population sizes (u, v, w) of the genotypes AA, Aa, and aa. Two important additional variables are the total population size $N=u+v+w$ and the gene frequency (i.e., the fraction of the gene A in the total gene pool) $p=(2u+v)/(2N)$. The latter, in particular, is the focus of attention in population genetics, and models formulated only in terms of this one variable are of special interest.

For a population with separate generations (and under the usual assumption of random mating), the two variables N and p completely determine the state immediately after the creation of a new generation, namely

$$(u, v, w)=(Np^2, 2Np(1-p), N(1-p)^2).$$

Simple assumptions concerning the nature of the selection process lead to a formula of the form

$$p_1=g_0(p)$$

for the gene frequency of the next generation, where

$$g_0(p)=\frac{W_1p^2+W_2p(1-p)}{W_1p^2+2W_2p(1-p)+W_3(1-p)^2}, \tag{1}$$

* Received by the editors April 13, 1987; accepted for publication (in revised form) March 7, 1988.
† Matematisk Institut, Århus Universitet, Ny Munkegade, Bygning 530, DK-8000 Århus C, Denmark.

and the positive numbers W_i, $i = 1, 2, 3$ can be interpreted as "fitnesses" of the three genotypes (they can arise in different ways in different models).

When the assumption is added that the population is distributed over a region in space, models can be formulated in terms of variables u, v, w, and N with similar interpretations as above, except that they are functions of space and time, and should be thought of as spatial densities. p is also a function of space and time with values that are real numbers between 0 and 1.

The papers by Felsenstein [9] and Weinberger [29] and the book by Fife [11] give impressions of the many different deterministic models that have been proposed for the description of the combined processes of selection, population size dynamics, and migration, as well as of the mathematical results obtained.

Whether or not the dynamics of a spatially distributed population with separate generations can be described in terms of the sole variable p depends on the assumptions concerning the interplay between the processes of selection, population regulation, and migration. If each generation is subject first to selection, then to population regulation (in the form that population density is set to a global constant without change in the local genotype proportions), and finally to migration, then the gene frequency distribution p_{n+1} of the offspring generation is derived from the gene frequency distribution p_n of the parent generation by

$$p_{n+1} = Qp_n, \tag{2}$$

where Q is a nonlinear integral operator of the form

$$Qp(x) = \int k(x, y) g(p(y))\, dy. \tag{3}$$

Here g denotes the function g_0 given by (1), and the kernel k describes the migration process (cf. Weinberger [29, formulae (2.1)-(2.4)]). If, on the other hand, regulation acts first, followed by selection and migration, then the gene frequency distribution p_{n+1} is given by

$$p_{n+1} = Rp_n, \tag{4}$$

where

$$Rp(x) = \frac{\int k(x, y) g(p(y))\, W(p(y))\, dy}{\int k(x, y)\, W(p(y))\, dy}. \tag{5}$$

Here, again, g is the function given by (1), and $W(p)$ is the denominator in (1):

$$W(p) = W_1 p^2 + 2W_2 p(1-p) + W_3(1-p)^2 \tag{6}$$

(cf. Weinberger [29, formula (2.12)]). It seems that, roughly speaking, these two situations are the only ones for which a model with one-dimensional state space is appropriate.

In the derivation of (2) and (4), the assumption of nonoverlapping generations is crucial. The standard model in the absence of this assumption is a differential equations model of the type considered by Fisher [13] and by Kolmogorov, Petrovskii, and Piskunov [19] in 1937:

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial x}\left(a \frac{\partial p}{\partial x}\right) + f(p) \tag{7}$$

(or its equivalent in higher-dimensional space). This model can be justified if the processes of selection and migration are slow when measured in terms of the life span

of individuals (cf., for instance, Nagylaki [23] and Weinberger [28, §8]). The selection process is slow if W_1, W_2, and W_3 are of the same order of magnitude, say $\bar{W}$. Then $W(p) \approx \bar{W}$ for $0 \leqq p \leqq 1$, and $g_0(p) - p \approx f_0(p)$, where

$$f_0(p) = p(1-p)(d_1 p - d_3(1-p)) \tag{8}$$

with

$$d_1 = (W_1 - W_2)/\bar{W}, \qquad d_3 = (W_3 - W_2)/\bar{W}.$$

Usually, the function f in (7) is assumed to be of the form (8) with (d_1, d_3) proportional to $(W_1 - W_2,\ W_3 - W_2)$. If the time scales of the processes of selection, migration, and reproduction are comparable, it seems that much more complicated models, such as models with age structure, are required.

2. Equilibrium distributions. Existence and stability. Known results. At the present stage of the presentation we will not be very specific about properties of the functions entering in the models (2), (4), and (7). As already indicated, we restrict ourselves to the case $n = 1$. We denote by Q and R, the operator given by (3) and (5), respectively where

p is a function from $\mathbf{R}$ to $[0, 1]$;

$g : [0, 1] \to [0, 1]$ is continuously differentiable and monotone;

$g(0) = 0$, $g(1) = 1$;

$W(p)$ is given by (6);

k is nonnegative and satisfies $\int k(x, y)\, dy = 1$ for all x.

Similarly, in order to talk conveniently about (7), we introduce the nonlinear differential operator A defined by

$$A(p) = \frac{\partial}{\partial x}\left(a \frac{\partial p}{\partial x}\right) + f(p),$$

where

the function $a : \mathbf{R} \to \mathbf{R}$ is strictly positive;

$f : [0, 1] \to \mathbf{R}$ is continuously differentiable; and

$f(0) = f(1) = 0$.

Depending on the model considered, an *equilibrium solution* is a function $p : \mathbf{R} \to [0, 1]$ satisfying $Qp = p$, $Rp = p$, or $Ap = 0$, respectively.

The constant functions $p(x) = 0$ and $p(x) = 1$ are trivial equilibria. Other obvious equilibria are constant functions $p(x) = \bar{p}$, where $\bar{p}$ is such that $g(\bar{p}) = \bar{p}$, or $f(\bar{p}) = 0$ respectively.

Less obvious, and mathematically and genetically more interesting, are nonconstant equilibrium solutions, so-called *clines*.

The existence of clines and their stability has been discussed by a large number of authors. It is a typical feature of these papers that stability rests on an assumption about spatial inhomogeneity (in this connection a subdivision of space into a finite or infinite collection of separate habitats is an extreme case of spatial inhomogeneity). See, e.g., Karlin and McGregor [16] for the case of several discrete habitats; Downham and Shah [7], Diekmann [6], and Lui [20] for the discrete time model (2); and Slatkin [27], Conley [4], Fleming [14], May, Endler, and McMurtrie [21], Nagylaki [23], Fife and Peletier [12], and Keller [18] for the diffusion equation model (7).

It is much less clear what to expect in homogeneous space, i.e., in the situation where the region considered is all of $\mathbf{R}$, and the operator Q, R, or A commutes with

translations. For the operators Q and R, this assumption means that W_1, W_2, and W_3 (and hence g) do not depend explicitly on x, and that k is of the form $k(x, y) = h(x-y)$. For the operator A it means that a is constant and f does not depend explicitly on x.

For (7) the situation is well understood. If $f(p)$ is of the form (8), and if $W_2 \geqq \min(W_1, W_3)$, then every equilibrium is constant. If $W_2 < \min(W_1, W_3)$ (the heterozygote is inferior), then there exist nonconstant equilibria, and every nonconstant equilibrium p_0 has one of the following properties:

(1) p_0 is periodic;

(2) p_0 has a single maximum, and $p_0(x) \to 0$ for $|x| \to \infty$, i.e., p_0 represents a "pocket" containing the (superior) gene A surrounded by a population in which the gene a is predominant;

(3) p_0 has a single minimum, and $p_0(x) \to 1$ for $|x| \to \infty$, i.e., p_0 represents a "pocket" containing the (inferior) gene a surrounded by a population in which the gene A is predominant;

(4) p_0 is a monotone cline, and $p_0(x) \to 0$ for $x \to -\infty$, $p_0(x) \to 1$ for $x \to \infty$ or conversely.

For a given set of W's, exactly one of the cases (2)-(4) occurs, depending on whether $W_1 > W_3$, $W_1 < W_3$, or $W_1 = W_3$. Furthermore, a nonconstant equilibrium is unstable unless it is monotone. See, for instance, McKean [22], Bazykin [3], Fife [10], Fife [11, §4.3], Razževaĭkin [24], Rosen [25], and Hagan [15].

Much less is known about the time evolution given by (2) or (4). It can be deduced from Theorem 6.5 in Weinberger [29] that if $g(p)$ is of the form (1), and if $W_2 \geqq \min(W_1, W_3)$, then every equilibrium is constant. In a forthcoming paper we will prove that if the heterozygote is inferior, then there exist nonconstant equilibria. It is not known, however, whether in the inferior heterozygote case, every nonconstant equilibrium has one of the properties (1)-(4), but we conjecture that this is the case. The purpose of the present paper is to prove an instability result analogous to the one mentioned above, viz. that a nonconstant equilibrium for (2) or (4) is unstable unless it is monotone. Since we must assume that the equilibrium considered has one of the properties (1)-(3), we do not quite succeed in this purpose.

3. The role of spatial homogeneity. Heuristics. A number of the assertions in this section should be taken with a grain of salt, for in order to bring out the basic ideas we will not worry about specifying precise conditions under which all steps are valid.

We consider the case of homogeneous space and assume that p_0 is an equilibrium solution to (2). For $r \in \mathbf{R}$ we let p_r denote the translate of p_0 defined by

$$p_r(x) = p_0(x+r). \tag{9}$$

Then, by the assumption of homogeneity, p_r is also an equilibrium solution, i.e.,

$$Qp_r = p_r \quad \text{for all real } r. \tag{10}$$

Consequently, *if* we work in a space F of functions having a differentiable structure, say F is a Banach space or a Banach manifold, and *if* the curve $r \mapsto p_r$ in F is differentiable at 0, and *if* Q is differentiable at p_0, then it follows by differentiation in (10) that

$$Q'p' = p'. \tag{11}$$

Here Q' denotes the derivative of Q at p_0 and p' denotes the derivative of p_r at the point $r = 0$. Now, with p_r defined by (9), it is true in "most" function spaces F that if the derivative p' exists, it is equal to the usual derivative p_0' of p_0 with respect

to x (for this to be true it is sufficient that the natural injection of F into the space $\mathcal{D}'$ of distributions is continuous and that p_0' is taken in the distribution sense). Also, in "most" function spaces it is true that if the derivative Q' exists at p, it is given by the linear integral operator

$$Q'(p_0)u(x) = \int h(x-y)g'(p_0(y))u(y)\,dy.$$

Since g is monotone, g' is nonnegative, so $Q'(p_0)$ is an integral operator with nonnegative kernel, and by (11) the function p_0' is an eigenfunction corresponding to the eigenvalue 1. Finally we note that "a lot of" integral operators with nonnegative kernal have the "Perron-Frobenius property," i.e., there exists a nonnegative eigenfunction, and its corresponding eigenvalue is simple and numerically maximal, i.e., it is equal to the spectral radius of the operator.

Clearly, if the preceding results apply, then an equilibrium solution p_0 to (2) that is not monotone must be unstable. This is true because p_0' is an eigenfunction for $Q'(p_0)$, and the Perron-Frobenius property implies that the corresponding eigenvalue, which is 1, is not the largest eigenvalue of $Q'(p_0)$ since the eigenfunction p_0' takes both positive and negative values. But if $Q'(p_0)$ has an eigenvalue larger than 1, then p_0 is an unstable equilibrium.

Next, consider an equilibrium solution of (4), i.e.,

$$p_0 = Rp_0 = \frac{Mp_0}{Np_0},$$

where

$$Mp(x) = \int h(x-y)g(p(y))W(p(y))\,dy, \tag{12}$$

$$Np(x) = \int h(x-y)W(p(y))\,dy. \tag{13}$$

As above, it follows from the assumption of spatial homogeneity that p_0' is an eigenfunction of the derivative $R'(p_0)$ corresponding to the eigenvalue 1, and, as above, under mild regularity assumptions, the operator $R'(p_0)$ is a linear integral operator with kernel

$$\begin{aligned} K(x,y) &= h(x-y)[Np_0(x)]^{-2}[Np_0(x)(gW)'(p_0(y)) - Mp_0(x)W'(p_0(y))] \\ &= h(x-y)[Np_0(x)]^{-1}[(gW)'(p_0(y)) - p_0(x)W'(p_0(y))]. \end{aligned} \tag{14}$$

Since $0 \leqq p_0(x) \leqq 1$ for all x, the kernel K is nonnegative if

$$(gW)'(p) \geqq \max(0, W'(p)) \quad \text{on } [0,1]. \tag{15}$$

Precisely as above we see that if $R'(p_0)$ has the Perron-Frobenius property, then p_0 cannot be stable unless it is monotone.

Finally, if p_0 is an equilibrium solution of

$$\frac{\partial p}{\partial t} = Ap, \tag{16}$$

then it follows from the translation invariance of the equation that the derivative p_0' is a solution to the linear differential equation

$$a\frac{d^2u}{dx^2} + f'(p_0(x))u = 0.$$

For "a lot of" second-order differential operators of the form

$$Lu = a\frac{d^2u}{dx^2} + qu$$

with positive a there is a positivity result related to the Sturm oscillation theorems that, in this connection, can replace the Perron-Frobenius property. The result in question states that if the upper part of the spectrum of L is discrete, then L has a positive eigenfunction, and the corresponding eigenvalue is simple and is the largest point in the spectrum of L.

Again, if the preceding results apply, then an equilibrium solution p_0 to (16) that is not monotone must be unstable. To see this, observe that since p_0' is an eigenfunction for L, and since it takes both positive and negative values, by the above-mentioned positivity result the corresponding eigenvalue, which is zero, is not the largest eigenvalue of L. But if L has an eigenvalue larger than zero, then p_0 is unstable. See Remark 2 below for a strengthening of this statement.

4. Formulation of assumptions and theorems.

Assumptions A. The function $g:[0,1]\to[0,1]$ satisfies the following:

(i) $g(0)=0$, $g(1)=1$;
(ii) g is differentiable with Hölder continuous derivative;
(iii) $g'(p)>0$ on $]0,1[$;
(iv) $g'(0)<1$.

Assumptions B. The function $h:\mathbf{R}\to\mathbf{R}$ satisfies the following:

(i) $h(x)\geqq 0$ on $\mathbf{R}$;
(ii) h is continuously differentiable;
(iii) $\int h(x)\,dx = 1$;
(iv) $\int |h'(x)|\,dx < \infty$;
(v) the closed additive subsemigroup of $\mathbf{R}$ generated by supp (h) is all of $\mathbf{R}$.

Assumptions C. The function $W:[0,1]\to\mathbf{R}$ satisfies the following:

(i) $W(p)>0$ on $[0,1]$;
(ii) W is differentiable with Hölder continuous derivative;
(iii) $(gW)'(p)\geqq \max(0, W'(p))$ for $p\in]0,1[$.

DEFINITION. By P we denote the subset of L^∞ defined by

$$P=\{p\in L^\infty : 0\leqq p(x)\leqq 1 \text{ a.e.}\}.$$

THEOREM 1. *Let Assumptions* A *and* B *be satisfied, and let Q be defined by* (*cf.* (3))

$$Qp(x) = \int h(x-y)g(p(y))\,dy. \tag{17}$$

Let p_0 be a solution to the equation $Qp=p$, assume that p_0 is not constant, and that either

$$p_0(x)\to 0 \quad \textit{for } |x|\to\infty, \quad \textit{or}$$

$$p_0 \quad \textit{is periodic.}$$

Then p_0 is an unstable equilibrium for (2) *considered as a discrete dynamical system in P.*

THEOREM 2. *Let Assumptions* A, B, *and* C *be satisfied, and let R be defined by* (*cf.* (5))

$$Rp(x) = \frac{\int h(x-y)g(p(y))W(p(y))\,dy}{\int h(x-y)W(p(y))\,dy}. \tag{18}$$

Let p_0 be a solution to the equation $Rp = p$, assume that p_0 is not constant, and that either

$$p_0(x) \to 0 \quad \textit{for } |x| \to \infty, \quad \textit{or}$$

$$p_0 \quad \textit{is periodic.}$$

Then p_0 is an unstable equilibrium for (4) *considered as a discrete dynamical system in P.*

Remark 1. For the sake of wider applicability, the results have been given a formulation independent of population genetics. Let us briefly describe their application to the models of population genetics.

If g is given by (1), then Assumptions A(i)-(iii) hold for all positive W_1, W_2, W_3, and Assumption A(iv) holds if and only if the heterozygote is inferior. If W is given by (6), then Assumptions C(i)-(ii) hold for all positive W_1, W_2, W_3, and Assumption C(iii) holds if and only if $W_2 \leqq 2 \min(W_1, W_3)$. Thus, if the heterozygote is inferior, Assumptions A and C hold. In this case it is also true that $g'(1) < 1$, and then the conclusions of Theorems 1 and 2 hold also for equilibria p_0 that satisfy $p_0(x) \to 1$ for $|x| \to \infty$.

Let us also note that (as mentioned above) if the heterozygote is not inferior, then all equilibria are constant.

Remark 2. As mentioned in § 2, it is known that nonmonotone equilibrium solutions of the nonlinear diffusion equation (16) are unstable. Actually, they have the so-called "hair-trigger" instability property (cf. Aronson and Weinberger [1], [2]), which is much stronger than that which can be proved along the lines considered in this paper. Instability results of hair-trigger type have also been proved for the unstable *constant* equilibria of (2) and (4) (cf. Weinberger [29]), and although we have not been able to prove it, we conjecture that similar results are valid for nonmonotone clines.

5. Proofs of the theorems.

Proof of Theorem 1. The proof is divided into several steps.

(i) p_0 does not assume the values 0 or 1.

Let us prove that p_0 does not assume the value 0. That it does not assume the value 1 follows by means of the transformation $p \mapsto 1 - p$ and Assumption B(iii).

Since p_0 is not constant, it is not identically 0, and $g(p_0)$ is different from 0 where p_0 is. Since both h and $g(p_0)$ are nonnegative, it follows from (17) that the support of p_0 contains the set $\operatorname{supp}(p_0) + \operatorname{supp}(h)$.

But then we first deduce from Assumption B(v) that $\operatorname{supp}(p_0)$ is all of $\mathbf{R}$, and, next, that $p_0(x) > 0$ for all x.

(ii) p_0 is twice continuously differentiable.

Since $p_0 = Qp_0$, it follows from Assumptions B(iv) and A(ii) that p_0 is differentiable with

$$p_0'(x) = \int h'(x-y) g(p_0(y))\, dy$$

$$= \int h(x-y) g'(p_0(y)) p_0'(y)\, dy.$$

Here, h can again be differentiated, and the claim follows.

(iii) Q is differentiable.

In fact, Q is differentiable at any $p \in P$ with

$$Q'(p)u(x) = \int h(x-y) g'(p(y)) u(y)\, dy. \tag{19}$$

It is clear that the linear map given by (19) is bounded in L^∞. To see that it is the derivative of Q note that

$$Q(p+u)-Q(p)=\int h(x-y)[g(p(y)+u(y))-g(p(y))]\,dy$$

$$=\int dy\int_0^1 h(x-y)g'(p(y)+tu(y))u(y)\,dt,$$

and use the uniform continuity of g'. See also the remark following the proof.

(iv) Q' is Hölder continuous.

This is an immediate consequence of the Hölder continuity of g'. We note in passing that it is essential that the topology in P be derived from the L^∞-norm and not from an L^r-norm with $r<\infty$.

In the rest of the proof we consider separately the case $p_0(x)\to 0$ at ∞ ((v)-(viii) below) and the case of periodic p_0 ((x) below).

First we discuss the case $p_0(x)\to 0$ at ∞.

(v) The linear map $K=Q'(p_0)$ has the Perron-Frobenius property.

More precisely, the spectral radius r of K is an eigenvalue of multiplicity 1, and it has a nonnegative eigenvector.

We will deduce this from the following result (cf. Schaefer [26, Thm. 3.2, p. 270]).

Let E be a Banach lattice and K an irreducible positive continuous linear mapping in E with spectral radius r. If r is a pole of the resolvent of K, then r has algebraic multiplicity 1, *and there exists a nonnegative eigenvector.*

The positivity of K follows from Assumption A(iv) and its irreducibility from Assumption B(v), so in order to establish (v) it remains to prove that r is a pole of the resolvent. As we shall see in (vii) below, this statement follows from step (vi).

(vi) There exist operators K_1 and K_2 in L^∞ such that $K=K_1+K_2$, $\|K_1\|<1$, K_2 is compact.

To prove this, we note that by Assumption A(iv) and the assumption that $p_0\to 0$ at ∞, there exist real numbers $c<1$ and C such that

$$g'(p_0(y))\leqq c \quad \text{for } |y|>C. \tag{20}$$

Then define the operators K_1 and K_2 by

$$K_1u(x)=\int_{|y|>C} h(x-y)g'(p_0(y))u(y)\,dy,$$

$$K_2u(x)=\int_{|y|\leqq C} h(x-y)g'(p_0(y))u(y)\,dy.$$

It is clear that $K=K_1+K_2$, and from Assumption B(iii) and (20) it follows that $\|K_1\|\leqq c<1$.

Let h_y denote the function $h_y(x)=h(x-y)$. Since h is uniformly continuous, the function $y\mapsto h_y$ from $\mathbf{R}$ to L^∞ is continuous, and hence the set $H=\{h_y:|y|\leqq C\}$ is a compact subset of L^∞.

By a theorem of Mazur (cf. Dunford and Schwartz [8, p. 416]), the closed convex hull $\hat{H}$ of $H\cup(-H)$ is also compact, and for any u in L^∞ we have $K_2u\in 2CM\|u\|\hat{H}$, where M denotes the maximum of g'.

The compactness of K_2 follows.

(vii) The spectral radius r of K is a pole of the resolvent of K.

Let $\sigma_e(K)$ denote the *essential spectrum* of the linear operator K, i.e., the set of complex numbers λ such that $\lambda I - K$ is *not* a semi-Fredholm operator (cf. Kato [17, p. 243]).

Then $\sigma_e(K_1)$ is contained in the spectrum of K_1, and hence in the closed disc of radius $\|K_1\|$. Since K_2 is compact, it follows from the Stability Theorem 5.35 in Kato [17, p. 244] that $\sigma_e(K)$ is contained in the same disc. Since K is bounded, it follows from Kato [17, p. 243] that, except for a discrete set of eigenvalues of finite algebraic multiplicity, the complement of this disc belongs to the resolvent set of K.

Now observe that 1 is an eigenvalue of K (with the eigenvector p_0'), and that, consequently, $r \geqq 1 > \|K_1\|$.

It is well known (cf., for instance, Schaefer [26, p. 264]), that r is a point in the spectrum of K, and the facts collected above show that it is an isolated eigenvalue of finite algebraic multiplicity. Thus (cf., for instance, Kato [17, III.6.5]), it is a pole of the resolvent of K.

(viii) p_0 is unstable.

By (v), r has multiplicity 1, and there exists a positive eigenvector corresponding to r. Since p_0' takes both positive and negative values, it cannot be an eigenvector corresponding to the eigenvalue r. On the other hand, p_0' *is* an eigenvector for K corresponding to the eigenvalue 1, and it follows that $r > 1$.

The instability of p_0 is now a consequence of (ix) below. The required regularity of Q was established in (iv). Also see the remark following the proof.

(ix) A discrete instability theorem.

Let U be a neighborhood of 0 in a Banach space F, let $f: U \to F$ be of class $C^{1+\alpha}$ for some α with $0 < \alpha < 1$, and assume that 0 is a fixed point of f. If the spectral radius r of the differential $K = df$ of f at 0 satisfies $r > 1$, then 0 is an unstable equilibrium of f considered as a discrete dynamical system in U.

This result is a straightforward adaptation to the discrete time situation of Theorem 2.3 in Daleckiĭ and Kreĭn [5, Chap. VII].

(x) The case of periodic p_0.

Let l denote the period of p_0, define L^∞_{per} to be the subspace of L^∞ consisting of functions of period l, and put $P_{\text{per}} = P \cap L^\infty_{\text{per}}$. Then it is clear that P_{per} is invariant under Q, and in order to prove that p_0 is unstable in P it is sufficient to prove that it is unstable in P_{per}.

Define

$$h_{\text{per}}(x) = \sum_{n=-\infty}^{\infty} h(x - nl),$$

then h_{per} is periodic with period l, and it follows from Assumptions B that

$h_{\text{per}}(x) \geqq 0$ for $x \in \mathbf{R}$;

$\int_0^l h_{\text{per}}(x)\, dx = 1$;

h_{per} is absolutely continuous; and

$\int_0^l |h_{\text{per}}'(x)|\, dx < \infty$.

Clearly, the restriction of Q to P_{per} is given by

$$Qp(x) = \int_0^l h_{\text{per}}(x - y) g(p(y))\, dy,$$

and, similarly, the restriction of the differential $K = Q'(p_0)$ to L^∞_{per} is given by

$$Ku(x) = \int_0^l h_{\text{per}}(x - y) g'(p_0(y)) u(y)\, dy. \tag{21}$$

The proof now proceeds as above, but it is simplified by the fact that the operator K defined on L^∞_{per} by (21) is compact. □

Remark. Strictly speaking, the statements about differentiability of Q and the use of the instability theorem (ix) require P to be an open subset of L^∞, and this is clearly not the case. This difficulty is most easily overcome by extending g to an *open* interval containing $[0, 1]$.

To prove instability of p_0 *inside* P we must restrict ourselves to the consideration of initial values in P. The initial values which, in the proof of (ix), are used to prove the instability, are of the form εu, where u is an eigenvector for K corresponding to the eigenvalue r. In the application to the present situation u is nonnegative, so that, fortunately, $p_0 + \varepsilon u \in P$ for ε small and positive.

Proof of Theorem 2. The proof is quite similar to the proof of Theorem 1, and many details will be omitted.

Note that $W(p)$ is bounded away from 0 for all $p \in P$ in view of Assumption C(i), and hence, so is the denominator in (18).

Define the operators M and N as in (12) and (13).

(i) First, we prove that p_0 does not assume the values 0 and 1, and that it is twice continuously differentiable.

Also, R is differentiable,

$$R'(p_0)u(x) = \int K(x, y)u(y)\, dy,$$

where the kernel K is given by (14), and, consequently, R' is Hölder continuous.

It follows from Assumptions C(i), (iii) that K is nonnegative. The proof of the irreducibility of $R'(p_0)$ uses the strict inequalities in Assumption C(iii), the fact that p_0 does not assume the values 0 and 1, and Assumption B(v). In the periodic case $R'(p_0)$ is compact, just as $Q'(p_0)$ was, so only the case $p_0 \to 0$ at ∞ requires further consideration.

(ii) As in (v) in the proof of Theorem 1 we will show that there exist operators K_1 and K_2 in L^∞ such that $R'(p_0) = K_1 + K_2$, $\|K_1\| < 1$, K_2 is compact.

First note that $W(p_0(y)) \to W(0)$ for $|y| \to \infty$, and hence, by dominated convergence,

$$Np_0(x) = \int h(y)\, W(p_0(x-y))\, dy \to \int h(y)\, W(0)\, dy = W(0) \quad \text{for } |x| \to \infty.$$

Now define

$$H_1(x, y) = h(x-y)[Np_0(x)]^{-1} W(p_0(y))g'(p_0(y)),$$
$$H_2(x, y) = h(x-y)[Np_0(x)]^{-1} g(p_0(y))\, W'(p_0(y)),$$
$$H_3(x, y) = -h(x-y)[Np_0(x)]^{-1} p_0(x)\, W'(p_0(y)).$$

If we choose c such that $g'(0) < c < 1$ (cf. Assumption A(iv)), then there exists a real number C such that

$$|x| > C \text{ and } |y| > C \quad \Rightarrow \quad [Np_0(x)]^{-1} W(p_0(y))g'(p_0(y)) \leqq c. \tag{22}$$

Let χ denote the characteristic function of the interval $[-C, C]$, define

$$H_{11}(x, y) = H_1(x, y)(1-\chi(x))(1-\chi(y)),$$
$$H_{12}(x, y) = H_1(x, y)(1-\chi(x))\chi(y),$$
$$H_{13}(x, y) = H_1(x, y)\chi(x),$$

and note that the kernel K can be written

$$K = H_{11} + H_{12} + H_{13} + H_2 + H_3.$$

It follows from (22) that the integral operator K_{11} with kernel H_{11} has operator norm at most equal to c, and, as in (vi) in the proof of Theorem 1, that the integral operator K_{12} with kernel H_{12} is compact.

Since $g(p_0(y)) \to 0$ at ∞, the operator K_2 with kernel H_2 is the uniform limit of the compact operators K_{2n} defined by

$$K_{2n}u(x) = \int_{-n}^{n} H_2(x, y)u(y)\, dy,$$

and consequently K_2 is compact.

The operators K_{13} and K_3 with kernels H_{13} (respectively H_3), are the adjoints in L^∞ of integral operators in L^1 with kernels that are the transposes of H_{13}, (respectively H_3). The proofs of the compactness of these L^1-operators are identical with the proofs of the compactness of K_{12} and K_2. But when the L^1-operators are compact, then so are their adjoints K_{13} and K_3.

We have now shown that

$$R'(p_0) = K_{11} + K_{12} + K_{13} + K_2 + K_3$$

with $\|K_{11}\| \leqq c < 1$ and K_{12}, K_{13}, K_2, and K_3 compact, and from here the proof proceeds as for Theorem 1. □

Acknowledgment. The author expresses his gratitude toward Tim Prout, who originally called the topic of cline stability to his attention.

REFERENCES

[1] D. G. ARONSON AND H. F. WEINBERGER, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics, J. A. Goldstein, ed., Springer-Verlag, Berlin, New York, 1975, pp. 5-49.

[2] ———, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. in Math., 30 (1978), pp. 33-76.

[3] A. D. BAZYKIN, *On a property of disruptive selection in a spatially distributed population.* I-III, Problemy Evolyucii, 2 (1972), pp. 219-223, 224-227, 228-232. (In Russian.)

[4] C. CONLEY, *An application of Wazewski's method to a non-linear boundary value problem which arises in population genetics*, J. Math. Biol., 2 (1975), pp. 241-249.

[5] JU. L. DALECKIĬ AND M. G. KREĬN, *Stability of Solutions of Differential Equations in Banach Space*, American Mathematical Society, Providence, RI, 1974.

[6] O. DIEKMANN, *Clines in a discrete time model in population genetics*, in Biological Growth and Spread, W. Jäger, H. Rost, and P. Tautu, eds., Springer-Verlag, Berlin, New York, 1980, pp. 267-275.

[7] D. Y. DOWNHAM AND S. M. M. SHAH, *A sufficiency condition for the stability of an equilibrium*, Adv. in Appl. Probab., 8 (1976), pp. 4-7.

[8] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part* I: *General Theory*, Interscience, New York, 1958.

[9] J. FELSENSTEIN, *The theoretical population genetics of variable selection and migration*, Ann. Rev. Genetics, 10 (1976), pp. 253-280.

[10] P. C. FIFE, *Stationary patterns for reaction-diffusion equations*, in Nonlinear Diffusion, W. E. Fitzgibbon (III) and H. F. Walker, eds., Pitman, London, 1977, pp. 81-121.

[11] ———, *Mathematical Aspects of Reacting and Diffusing Systems*, Springer-Verlag, Berlin, New York, 1979.

[12] P. C. FIFE AND L. A. PELETIER, *Nonlinear diffusion in population genetics*, Arch. Rational Mech. Anal., 64 (1977), pp. 93-109.

[13] R. A. FISHER, *The wave of advance of advantageous genes*, Ann. Eugenics, 7 (1936-37), pp. 355-369.
[14] W. H. FLEMING, *A selection-migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219-233.
[15] P. S. HAGAN, *The instability of nonmonotonic wave solutions of parabolic equations*, Stud. Appl. Math., 64 (1981), pp. 57-88.
[16] S. KARLIN AND J. MCGREGOR, *Application of method of small parameters to multi-niche population genetic models*, Theoret. Population Biol., 3 (1972), pp. 186-209.
[17] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1966.
[18] J. B. KELLER, *Genetic variability due to geographical inhomogeneity*, J. Math. Biol., 20 (1984), pp. 223-230.
[19] A. N. KOLMOGOROV, I. G. PETROVSKIĬ, AND N. S. PISKUNOV, *Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application a un problème biologique*, Bull. Univ. Etat Moscou (Sér. Int.), Sec. A1, Fasc. 6 (1937), pp. 1-25.
[20] R. LUI, *A nonlinear integral operator arising from a model in population genetics. IV. Clines*, SIAM J. Math. Anal., 17 (1986), pp. 152-168.
[21] R. M. MAY, J. A. ENDLER, AND R. MCMURTRIE, *Gene frequency clines in the presence of selection opposed by gene flow*, Amer. Naturalist, 109 (1975), pp. 659-676.
[22] H. P. MCKEAN, *Nagumo's equation*, Adv. in Math., 4 (1970), pp. 209-223.
[23] T. NAGYLAKI, *Conditions for the existence of clines*, Genetics, 80 (1975), pp. 595-615.
[24] V. N. RAZŽEVAĬKIN, *The instability of stationary inhomogeneous solutions of the Cauchy problem for the quasilinear parabolic equation and its ecological applications*, Zh. Vychisl. Mat. i Mat. Fiz., 20 (1980), pp. 1328-1333. (In Russian.) U.S.S.R. Comput. Math. and Math. Phys., 20, No. 5 (1980), pp. 235-240. (In English.)
[25] G. ROSEN, *On the Fisher and the cubic-polynomial equations for the propagation of species properties*, Bull. Math. Biol., 42 (1980), pp. 95-106.
[26] H. H. SCHAEFER, *Topological Vector Spaces*, Macmillan, New York, 1966.
[27] M. SLATKIN, *Gene flow and selection in a cline*, Genetics, 75 (1973), pp. 733-756.
[28] H. F. WEINBERGER, *Asymptotic behavior of a model in population genetics*, in Nonlinear Partial Differential Equations and Their Applications, J. M. Chadam, ed., Springer-Verlag, Berlin, New York, 1978, pp. 47-96.
[29] ———, *Long-time behavior of a class of biological models*, SIAM J. Math. Anal., 13 (1982), pp. 353-396.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

012

EXISTENCE OF BEST PARAMETRIC INTERPOLATION BY CURVES*

K. SCHERER† AND P. W. SMITH‡

Abstract. Given n points in $\mathbb{R}^d$, find an acceptable interpolating curve. This problem is addressed in this paper. Analogues of the natural spline interpolators are examined, and the nonlinear problem that arises when the interpolation abscissae are treated as variables in the minimum seminorm problem is discussed in detail.

Key words. splines, interpolation, parametric curves

AMS(MOS) subject classifications. 41A15, 41A05

1. Introduction. The following problem is addressed. Given vectors $\mathbf{z}_1, \cdots, \mathbf{z}_n \in \mathbb{R}^d$, the d-dimensional Euclidean space, and given points

$$\mathbf{t}: a = t_1 < t_2 < \cdots < t_n = b, \tag{1}$$

we introduce the class

$$U_{k,\mathbf{t}}(\{\mathbf{z}_i\}_{i=1}^n) := \{\mathbf{f}(t) = (f_1(t), \cdots, f_d(t)): f_j \in L_2^k(a, b), \mathbf{f}(t_i) = \mathbf{z}_i, 1 \leqq i \leqq n\} \tag{2}$$

where $L_2^k(a, b)$ is the space of functions with $k-1$ absolutely continuous derivatives and kth derivative in $L_2(a, b)$. Hence $U_{k,\mathbf{t}}$ is the class of all smooth curves in $\mathbb{R}^d$ which pass through the points $\mathbf{z}_i \in \mathbb{R}^d$ at the time t_i, $1 \leqq i \leqq n$. Then we look for a best interpolating curve in the sense that

$$I = \inf_{\mathbf{t}} \inf \{|||\mathbf{f}^{(k)}|||: \mathbf{f}^{(k)} := (f_1^{(k)}, \cdots, f_d^{(k)}), \mathbf{f} \in U_{k,\mathbf{t}}(\{\mathbf{z}_i\}_{i=1}^n)\} \tag{3}$$

is attained, where

$$|||\mathbf{f}||| := \sqrt{\int_a^b \sum_{i=1}^d |f_i(t)|^2 \, dt}. \tag{4}$$

We are interested in existence and uniqueness of solutions of (3), in particular in determining optimal knots $\mathbf{t} = \{t_1, \cdots, t_n\}$. If we work with fixed knots, i.e., we take in (3) the infimum only with respect to $U_{k,\mathbf{t}}(\{\mathbf{z}_i\}_{i=1}^n)$, the problem reduces to the classical spline problem. The geometric motivation for (3) comes from the fact that curves may be parametrized differently, which influences the choice of the t_i. An interesting physical interpretation in case $k = 2$ is that a solution of (3) presents a trajectory with the least kinetic energy through prescribed points in $\mathbb{R}^d$ (see, e.g., Töpfer [3], Marin [2]). In the latter paper, existence and uniqueness of a solution of (3) has been shown in case $k = 2$, $d = 1$.

2. Existence. We assume from now on that $n > k$, since otherwise (3) would have a trivial solution in π_k, the class of curves with component functions being polynomials of degree $< k$. Next we introduce the following subclass of $U_{k,\mathbf{t}}$ (for data $\{\mathbf{w}_i\}_{i=1}^n$ with $\mathbf{w}_1 = 0$):

$$U_{k,\mathbf{t}}^0(\{\mathbf{w}_i\}_{i=2}^n) := \{\mathbf{g} \in U_{k,\mathbf{t}}(\{\mathbf{w}_i\}_{i=1}^n): \mathbf{g}^{(\nu)}(a) = 0, 1 \leqq \nu \leqq k-1, g(a) \equiv \mathbf{w}_1 = 0)\}. \tag{5}$$

With the help of this notion the problem can be reformulated.

* Received by the editors March 31, 1986; accepted for publication (in revised form) March 14, 1988.
† Institut für Angewandte Mathematik der Universität Bonn, Bonn, West Germany.
‡ IMSL, 2500 Park West Tower One, 2500 City West Boulevard, Houston, Texas 77042-3020.

LEMMA 1. *The following relation holds*:

$$I = \inf_{\mathbf{p}\in\pi_k} \{\inf_{\mathbf{t}} I_{\mathbf{t},\mathbf{p}}: p(a) = \mathbf{z}_1\} \tag{6}$$

where for fixed $\mathbf{t}$ *satisfying* (1)

$$I_{\mathbf{t},\mathbf{p}} := \inf\{|||\mathbf{g}^{(k)}|||: \mathbf{g}\in U^0_{k,\mathbf{t}}(\{\mathbf{z}_i - \mathbf{p}(t_i)\}_{i=2}^n)\}. \tag{7}$$

Furthermore, to any solution $\mathbf{t}^*$, $\mathbf{f}^*$ *of* (3) *there corresponds a solution* $\mathbf{t}^*$, $\mathbf{g}^*$, *and* $\mathbf{p}^*$ *of the right-hand side of* (6) *and conversely. In this case* $\mathbf{p}^*$ *is the Taylor polynomial* $\in \pi_k^*$ *of* $\mathbf{f}^*$ *at the point* a.

Proof. For each $\mathbf{f}\in U_{k,\mathbf{t}}$ we write

$$\mathbf{f}(t) = \sum_{\nu=0}^{k-1} \mathbf{f}^{(\nu)}(a)(t-a)^\nu/\nu! + \mathbf{g}(t) \equiv \mathbf{p}_f(t) + \mathbf{g}(t). \tag{8}$$

Then we have $\mathbf{f}^{(k)} = \mathbf{g}^{(k)}$, $\mathbf{p}_f \in \pi_k$ with $\mathbf{p}_f(a) = \mathbf{z}_1$ and $\mathbf{g}\in U^0_{k,\mathbf{t}}(\{\mathbf{z}_i - \mathbf{p}_f(t_i)\}_{i=2}^n)$ so that for any fixed set $\mathbf{t} = (t_1, \cdots, t_n)$ it follows that

$$\inf\{|||f^{(k)}|||: f\in U_{k,\mathbf{t}}(\{z_i\}_{i=1}^n)\} \geqq I_{\mathbf{t},\mathbf{p}} \geqq \inf_{\mathbf{p}\in\pi_k} \{\inf_{\mathbf{t}} I_{\mathbf{t},\mathbf{p}}: \mathbf{p}(a) = \mathbf{z}_1\}.$$

Taking the inf with respect to $\mathbf{t}$, we obtain "$\geqq$" in (6). On the other hand, for any fixed $\mathbf{p}\in\pi_k$ we have trivially

$$I \leqq \inf_{\mathbf{t}} \inf\{|||\mathbf{g}^{(k)}|||: \mathbf{g}\in U^0_{k,t}(\{\mathbf{z}_i - \mathbf{p}(t_i)\}_{i=2}^n)\}, \mathbf{g}(t_i) + \mathbf{p}(t_i) = \mathbf{z}_i, \mathbf{p}(a) = \mathbf{z}_1\},$$

which proves equality in (6) after taking the inf with respect to $\mathbf{p}$. Now if there were a solution $\mathbf{t}^*$, $\mathbf{f}^*$ with $\mathbf{f}^*$ attaining I, it could be decomposed as in (8) with $I = |||\mathbf{g}^{*(k)}|||$ so that $\mathbf{g}^*$, $\mathbf{p}^*$, $\mathbf{t}^*$ would form a solution triple of the right-hand side of (6). On the other hand, such a solution triple yields via $\mathbf{f}^* := \mathbf{g}^* + \mathbf{p}^*$ a solution pair $\mathbf{t}^*, \mathbf{f}^*$ attaining I. □

Now we investigate the infimum problem (7).

LEMMA 2. *For each fixed* $\mathbf{p}\in\pi_k$ *and* $\mathbf{t} = (t_1, \cdots, t_n)$ *satisfying* (1) *the infimum* $I_{\mathbf{t},\mathbf{p}}$ *is uniquely attained by a function* $\mathbf{g}^*\in U^0_{k,\mathbf{t}}(\{\mathbf{z}_i - \mathbf{p}(t_i)\}_{i=2}^n)$ *of the form*

$$g^{*(k)}(t) = \sum_{i=2}^n \boldsymbol{\alpha}_i \varphi_i(t), \qquad \varphi_i(t) := (t_i - t)_+^{k-1}/(k-1)!. \tag{9}$$

The vectors $\boldsymbol{\alpha}_i \in \mathbb{R}^d$ *are uniquely determined via the system*

$$\sum_{i=2}^n \boldsymbol{\alpha}_i G_{ij} = \mathbf{z}_j - \mathbf{p}(t_j), \qquad 2\leqq j\leqq n \tag{10}$$

where $G_{\mathbf{t}} = (G_{ij})$ *is the Gramian matrix*

$$G_{ij} := (\varphi_i, \varphi_j) = \int_a^b \varphi_i(t)\varphi_j(t)\,dt, \qquad 2\leqq i,j\leqq n. \tag{11}$$

The value of the infimum $I_{\mathbf{t},\mathbf{p}}$ *is given by*

$$I^2_{\mathbf{t},\mathbf{p}} = \langle G_{\mathbf{t}}\boldsymbol{\alpha}, \boldsymbol{\alpha}\rangle = \langle G_{\mathbf{t}}^{-1}(\mathbf{z} - p_{\mathbf{t}}), \mathbf{z} - p_{\mathbf{t}}\rangle. \tag{12}$$

Here the scalar product is to be understood as a sum of n *scalar products in* $\mathbb{R}^d$ *namely between the* n *components* $\mathbf{z} - p_{\mathbf{t}}$ *and* $G_{\mathbf{t}}^{(-1)}(\mathbf{z} - p_{\mathbf{t}})$ *where* $\mathbf{z} = \{\mathbf{z}_i\}_{i=1}^n$, $p_{\mathbf{t}} = \{\mathbf{p}(t_i)\}_{i=1}^n$ *and* $G_{\mathbf{t}}^{-1}$ *denote the inverse of* $G_{\mathbf{t}}$ *above. (The scalar products are formed with respect to the subsequences of* $\mathbf{z} - p_{\mathbf{t}}$, *which arise by taking in* $\mathbf{z}_i$, $\mathbf{p}(t_i)$ *their* $\mathbb{R}^d$*-components.)*

Proof. The principal observation is that for $\mathbf{g} \in U^0_{k,\mathbf{t}}(\{\mathbf{w}_j\}_{j=2}^n)$ the point functional $g(t_i)$, $2 \leqq i \leqq n$, can be written as

$$g(t_i) = \int_a^b \varphi_i(t)\mathbf{g}^{(k)}(t)\,dt. \tag{13}$$

This follows easily by repeated partial integration yielding

$$\int_a^b \frac{(t_j - t)_+^{k-1}}{(k-1)!}\mathbf{g}^{(k)}(t)\,dt = \int_a^b \frac{(t_j - t)_+^{k-2}}{(k-2)!}\mathbf{g}^{(k-1)}(t)\,dt$$

$$= \int_a^b (t_j - t)_+^0 \mathbf{g}'(t)\,dt = \int_a^{t_j} \mathbf{g}'(t)\,dt = \mathbf{g}(t_i) = \mathbf{w}_i.$$

After choosing a function $\mathbf{g}_0 \in U^0_{k,\mathbf{t}}(\{\mathbf{w}_j\}_{j=2}^0)$, we can now write

$$I_{\mathbf{t},\mathbf{p}} = \inf\{|||\mathbf{g}^{(k)} + \mathbf{g}_0^{(k)}|||: \mathbf{g} \in U^0_{k,\mathbf{t}}(\{0\}_{j=2}^n)\}$$

$$= \inf\left\{|||\mathbf{g}^{(k)} + \mathbf{g}_0^{(k)}|||: \int_a^b \mathbf{g}^{(k)}(t)\varphi_j(t)\,dt = 0,\ 2 \leqq j \leqq n\right\}.$$

Hence we conclude, in the standard manner for Hilbert spaces, that the solution $\mathbf{g}^*(k)$ must be in the $\mathbb{R}^d$-valued span of $\{\varphi_j\}_{j=2}^n$. This gives (9) and conditions $\mathbf{g}^*(t_j) = \mathbf{w}_j = \mathbf{z}_j - \mathbf{p}(t_j)$, $2 \leqq j \leqq n$, determine g^* uniquely via (10) in view of (13). Finally, (12) is obtained by direct calculation. □

Remark. The procedure of Lemma 2 is well known for the classical spline problem, where a relation of type (13) is usually employed with divided differences and B-splines. The point here—to be used later—is that the right-hand side of (10) has a much simpler form than would be the case for the classical spline approach. Information about the size of the eigenvalues of the matrix $G_{\mathbf{t}}$ gives the following.

LEMMA 3. *For a sequence* $\mathbf{t}$ *satisfying* (1) *set*

$$\delta := \min(t_{j+1} - t_j). \tag{14}$$

Then for any sequence $\{\beta_i\}_{i=2}^n$, $\beta_i \in \mathbb{R}$ *there holds*

$$C(b-a)\delta^{2k} \sum_{i=2}^n |\beta_i|^2 \leqq \sum_{i,j=2}^n G_{ij}\beta_i\beta_j \leqq n(b-a)^{2k-1} \sum_{i=2}^n |\beta_i|^2 \tag{15}$$

where C *is an absolute constant.*

Proof. We have by definition

$$\sum_{i,j=2}^n G_{ij}\beta_i\beta_j = \int_a^b \left|\sum_{i=2}^n \beta_i\varphi_i(t)\right|^2 dt \leqq \sum_{i=2}^n |\beta_i|^2 \int_a^b \sum_{i=2}^n |\varphi_i(t)|^2.$$

From this the inequality from above in (15) follows immediately in view of the (rough) estimate $|\varphi_i(t)|^2 \leqq (b-a)^{2k-2}$ for $t \in [a, b]$.

For the converse inequality we construct functions $\psi_i \in L_\infty^{(k)}(a, b)$ with

$$\delta_{ij} = \int_a^b \varphi_j(t)\psi_i^{(k)}(t)\,dt, \qquad 2 \leqq i,\ j \leqq n. \tag{16}$$

Relation (13) says that this is equivalent to

$$\delta_{ij} = \psi_i(t_j).$$

Hence the choice $\psi_i(t) = h((t-t_i)/\delta)$ where h is a fixed function $\in L_\infty^k(-\infty, \infty)$ with support $(0, 1)$ will achieve this. It follows from (16) that (setting $s = \sum_{j=2}^n \beta_j \varphi_j$)

$$|\beta_i|^2 = \left[\int_a^b s(t)\psi_i^{(k)}(t)\right]^2 \leqq \int_a^b |\psi_i^{(k)}|^2 \int_a^b |s|^2 \leqq C(t_{i+1}-t_{i-1})\delta^{-2k}\int_a^b |s|^2.$$

Summing over i yields the other inequality in (15). □

Remark. The lower bound quantitatively exhibits the ill conditioning of the truncated power basis.

In the next step we settle the question of the dependence of $I_{\mathbf{t},\mathbf{p}}$ of $\mathbf{p}$.

LEMMA 4. *For each* $\mathbf{t}$ *satisfying* (1) *there is a unique* $\mathbf{p}_t^*$ *satisfying*

$$I_{\mathbf{t}} \equiv \inf_{\mathbf{p}} I_{\mathbf{t},\mathbf{p}} = \sqrt{\langle G_{\mathbf{t}}^{-1}(\mathbf{z}-p_{\mathbf{t}}^*), \mathbf{z}-p_{\mathbf{t}}^*\rangle} = \sqrt{\langle G_{\mathbf{t}}^{-1}\mathbf{z}, \mathbf{z}\rangle - \langle G_{\mathbf{t}}^{-1}p_{\mathbf{t}}^*, p_{\mathbf{t}}^*\rangle}. \tag{17}$$

Proof. Since $G_{\mathbf{t}}$ is symmetric and positive definite, so is $G_{\mathbf{t}}^{-1}$. Hence we can define the inner product

$$\langle \mathbf{u}, \mathbf{v}\rangle_G := \langle G_{\mathbf{t}}^{-1}\mathbf{u}, \mathbf{v}\rangle \tag{18}$$

for vectors $\mathbf{u}, \mathbf{v}$ with $n-1$ components in $\mathbb{R}^d$. Then (17) reduces to a problem of best approximation with respect to the norm induced by (18), where $\mathbf{z}$ is approximated by the linear space $V \equiv \{\{\mathbf{p}(t_j)\}_{j=2}^n : \mathbf{p} = \text{polynomial curve of degree } < k\}$. Thus (17) has a unique solution characterized by the relation

$$\langle G_{\mathbf{t}}^{-1}(\mathbf{z}-p_{\mathbf{t}}^*), \mathbf{q}\rangle = 0, \qquad \mathbf{q} \in V$$

from which the second inequality in (17) follows. □

We now turn to the question of existence. We observe that

$$I = \inf\{I_{\mathbf{t}} : \mathbf{t} \text{ satisfies (1)}\}. \tag{19}$$

Indeed we obviously have $I_{\mathbf{t},\mathbf{p}} \geqq I_{\mathbf{t}}$ and so by Lemma 1

$$I \geqq \inf_{\mathbf{p}} \inf_{\mathbf{t}} I_{\mathbf{t}} = \inf_{\mathbf{t}} I_{\mathbf{t}}.$$

On the other hand, (6) of Lemma 1 yields for any $\mathbf{t}$ satisfying (1)

$$I \leqq \inf_{\mathbf{p}} I_{\mathbf{t},\mathbf{p}} = I_{\mathbf{t}}$$

so that (19) must hold.

Next we introduce a definition.

DEFINITION. The data $\{\mathbf{z}_i\}_{i=1}^n$ are called asymptotically polynomial of order k if there is a sequence of polynomial curves $\mathbf{p}_N(t) \in \pi_k$ and a sequence of nodes $\{t_i^{(N)}\}_{i=1}^n$ satisfying (1) such that

$$\mathbf{z}_i = \lim_{N\to\infty} \mathbf{p}_N(t_i^{(N)}), \qquad 1 \leqq i \leqq n. \tag{20}$$

THEOREM. *If the data* $\{\mathbf{z}_i\}_{i=1}^n$ *are not asymptotically polynomial of order* k *and if the data are "rough," i.e.,*

$$\mathbf{z}_{i+1} \neq \mathbf{z}_i, \qquad 1 \leqq i \leqq n-1, \tag{21}$$

there exists a solution of (3), *i.e., a sequence* $\mathbf{t}^* = \{t_i^*\}_{i=1}^n$ *of simple knots* (*satisfying* (1)) *and a polynomial curve* $\mathbf{p}^* \in \pi_k$ *such that* $I = I_{\mathbf{t}^*,\mathbf{p}^*}$ *and* $I_{\mathbf{t}^*,\mathbf{p}^*}$ *has a unique solution* $\mathbf{g}^*$ *according to Lemma* 2. *The solution* f^* *of* (3) *is then obtained as described in Lemma* 1.

Proof. According to (19) we can assume

$$I = \lim_{N\to\infty} I_{\mathbf{t}^{(N)}}$$

for a sequence $\mathbf{t}^{(N)}$ of knot sequences satisfying (1). Let p_N^* be the corresponding "best" polynomials in the sense of (17) in Lemma 4. We then show that the values of

these polynomial curves at $\mathbf{t}^{(N)}$ remain uniformly bounded with the respect to N, i.e., for the Euclidean norm in $\mathbb{R}^d$

$$\sup_{2\leqq j\leqq n} \| p_N^*(t_j^{(N)})\| \leqq M, \qquad N\to\infty. \tag{22}$$

To this end, note that in view of the fact that the smallest eigenvalue of $G_{\mathbf{t}}^{-1}$ is the reciprocal of the largest eigenvalue of $G_{\mathbf{t}}$, relation (15) allows us to conclude that (with notation G_N for $G_{\mathbf{t}^{(N)}}$ and $p_{N,\mathbf{t}}$ for $\{p_N^*(t_i^{(N)})\}_{i=1}^n$)

$$\begin{aligned} 2I^2 &\geqq I_{\mathbf{t}(N)}^2 = \langle G_N^{-1}(\mathbf{z}-p_{N,\mathbf{t}}), \mathbf{z}-p_{N,\mathbf{t}}\rangle \\ &\geqq n^{-1}(b-a)^{1-2k}|\mathbf{z}-p_{N,\mathbf{t}}|^2 \geqq n^{-1}(b-a)^{1-2k}[|p_{N,\mathbf{t}}|-|z|]^2 \end{aligned} \tag{23}$$

where $|\ |$ notes the Euclidean norm on $\mathbb{R}^{(n-1)d}$. From this (22) follows.

We cannot conclude at this point that the polynomials themselves are bounded (but see (30)). However, we can assume (eventually passing to a subsequence) that

$$\lim_{N\to\infty} p_N^*(t_j^{(N)}) = \mathbf{q}_j \in \mathbb{R}^d, \qquad 2\leqq j\leqq n. \tag{24}$$

Now suppose (by way of contradiction) that

$$\lim_{N\to\infty} t_i^{(N)} = t_i^* \in [a, b] \tag{25}$$

and that there are only $r<k$ distinct values among the t_i^*; thus

$$a = t_1^* = \cdots = t_{i_1}^* < t_{i_2}^* < \cdots < t_{i_r}^* = \cdots = t_n^* = b. \tag{26}$$

Next we observe that (10) can be written as

$$\mathbf{z}_j - \mathbf{p}_N^*(t_j^{(N)}) = (\mathbf{g}_N^{(k)}(t), \varphi_{j,N}(t)) = \left(\sum_{i=2}^n \boldsymbol{\alpha}_i^{(N)}\varphi_{i,N}, \varphi_{i,N}\right) \tag{27}$$

where $\varphi_{i,N}(t) := (t_i^{(N)}-t)_+^{k-1}/(k-1)!$ and $\mathbf{g}_N^{(k)}(t) := \sum_{i=2}^n \boldsymbol{\alpha}_i^{(N)}\varphi_{i,N}(t)$ satisfies

$$|||\mathbf{g}_N^{(k)}|||^2 = \langle G_{\mathbf{t}}\boldsymbol{\alpha}^{(N)}, \boldsymbol{\alpha}^{(N)}(t)\rangle \to I^2, \qquad N\to\infty. \tag{28}$$

Subtraction of equations (27) for j, j' with $i_\nu < j, j' \leqq i_{\nu+1}$ gives

$$\begin{aligned} \|\mathbf{z}_j - \mathbf{z}_{j'} - \mathbf{p}_N^*(t_j^{(N)}) + \mathbf{p}_N^*(t_{j'}^{(N)})\| &= \|(\mathbf{g}_N^{(k)}, \varphi_{j,N}-\varphi_{j',N})\| \\ &\leqq |||\mathbf{g}_N^{(k)}|||\left(\int_a^b |\varphi_{j,N}(t)-\varphi_{j',N}(t)|^2\,dt\right)^{1/2}. \end{aligned}$$

Passing to the limit we obtain, in view of (24), (28),

$$\mathbf{z}_j - \mathbf{q}_j = \mathbf{z}_{j'} - \mathbf{q}_{j'}, \qquad i_\nu < j, j' \leqq i_{\nu+1}. \tag{29}$$

Now let $\tilde{p}$ be the polynomial curve in $\pi_r \subseteq \pi_k$ interpolating the values $\{\mathbf{z}_{i_j} - \mathbf{q}_{i_j}\}_{j=1}^r$ at the points $\{t_{i_j}^*\}$. Then consider the polynomials

$$\tilde{\mathbf{p}}_N(t) := \mathbf{p}_N^*(t) - \tilde{p}(t)$$

and the data

$$\mathbf{w}_j^{(N)} := \tilde{\mathbf{p}}_N(t_j^{(N)}).$$

In view of

$$\mathbf{w}_j^{(N)} = p_N^*(t_j^{(N)}) - (q_j - z_j) + \tilde{p}(t_j^*) - \tilde{p}(t_j^{(N)})$$

and (24), (29), this implies

$$\lim_{N\to\infty} w_j^{(N)} = \mathbf{z}_j, \qquad 1\leqq j\leqq n,$$

which is by (20) a contradiction to our assumption on the data. Therefore, we must have $r \geqq k$ in (26). Then we can use Newton's formula with the knots $t_{i_1}^{(N)}, \cdots, t_{i_k}^{(N)}$ for the representation of the polynomials $p_N^*(t)$. In view of (24) and (26) we have (for a subsequence)

$$\lim_{N\to\infty} p_N^*(t) = p^*(t) \tag{30}$$

uniformly in t for some polynomial curve $p^*(t)$. But then (24) and (29) imply that all knots are simple since otherwise we would have a contradiction to (21). Therefore, the matrices $G_N^{-1} \equiv G_{\mathbf{t}(N)}^{-1}$ tend to a nonsingular Gramian matrix $(G^*)^{-1}$ where $G^* \equiv (G_{ij}^*)$:

$$G_{ij}^* = (\varphi_i^*, \varphi_j^*), \qquad \varphi_i^*(t) := (t_i^* - t)_+^{k-1}/(k-1)!.$$

Hence passing to limit $N \to \infty$ in (27), we see that there exist $\boldsymbol{\alpha}_i^* \in \mathbb{R}^d$

$$\mathbf{z}_j - \mathbf{p}^*(t_j^*) = \left(\sum_{i=2}^{n} \boldsymbol{\alpha}_i^* \varphi_i^*, \varphi_j^* \right) =: (\mathbf{g}^{*(k)}, \varphi_j^*).$$

Therefore, the spline curve $\mathbf{g}^*$ satisfies $\mathbf{g}^*(t_j^*) = \mathbf{z}_j - \mathbf{p}^*(t_j^*)$ and is determined by the condition in (5). In addition it follows from (28) that $\|\|g^*\|\| = I$. Hence $\mathbf{g}^* + \mathbf{p}^*$ is a solution of (3). □

The motivation for working in the above theorem with the additional hypothesis of nonasymptotic-polynomial data came from the fact that a simple sufficient condition for this would be that $I \equiv I(\mathbf{z})$ is continuous with respect to the data $\mathbf{z} = \{z_i\}_{i=1}^n$ and that $I(\mathbf{z}) > 0$. Namely, the continuity implies $I(\mathbf{z}) = 0$ in case the data are asymptotically polynomial (of order k). We can show that $I(\mathbf{z})$ is always upper semicontinuous, but unfortunately there exist asymptotically polynomial data for which $I(\mathbf{z}) > 0$. A simple example is given for $k = 3$ by the data (in $\mathbb{R}^2$)

$$\mathbf{z}_1 = (-1, 1), \quad \mathbf{z}_2 = (-1, 0), \quad \mathbf{z}_3 = (1, 0), \quad \mathbf{z}_4 = (1, 1). \tag{31}$$

Then we easily check that the parabola $(x(t), y(t))$ given by

$$x(t) = t, \qquad y(t) = 1 + N - Nt^2$$

passes through $\mathbf{z}_1, \mathbf{z}_4$ for $t = -1$ and $t = 1$, respectively, and $(\pm\sqrt{1+1/N}, 0)$ for $t = \pm\sqrt{1+1/N}$. Hence it approximates the data (31) arbitrarily close as $N \to \infty$.

We note that here the nodes coalesce in the limit, which is exactly what is avoided in the theorem by our additional hypotheses. In fact, we can see from subsequent considerations that we have continuity with respect to the data if they allow a solution in the sense of the theorem.

In order to show that $I(\mathbf{z}) > 0$ for the data (31) we use the classical approach via divided differences and B-splines mentioned above. According to it (cf. de Boor [1]) the solution for $I_{\mathbf{t}}$ with $\mathbf{t}$ satisfying (1) can be written as

$$\mathbf{f}^{*(k)} = \sum_{i=1}^{n-k} \boldsymbol{\beta}_i M_{i,k,2} \tag{32}$$

where the $M_{i,k,2}$ are the B-splines normalized via $M_{i,k,2}(x) = [(t_{i+k} - t_i)/k]^{1/2} M_{i,k}(x)$ and $\int M_{i,k} = 1$, and $\boldsymbol{\beta}_i \in \mathbb{R}^d$ are determined by

$$\sum_{i=1}^{n-k} \boldsymbol{\beta}_i (M_{i,k,2}, M_{j,k,2}) = Z_j \equiv [(t_{j+k} - t_j)/k]^{1/2} [t_j, \cdots, t_{j+k}]\mathbf{z} \cdot k!. \tag{33}$$

From this we derive the formula

$$I_{\mathbf{t}}^2 = \sum_{i,j=1}^{n-k} K_{ij}(Z_i, Z_j)$$

where the K_{ij} are elements of the inverse of the Gramian $\{(M_{i,k,2}, M_{j,k,2})\}_{i,j=1}^{n-k}$. According to the well-known Isomorphism Theorem of de Boor [1], we know there is a constant D_k depending only on k such that

$$D_k^{-2}\sum_i |\alpha_i|^2 \leqq \int \left|\sum_i \alpha_i M_{i,k,2}(x)\right|^2 dx \leqq \sum_i |\alpha_i|^2 \tag{34}$$

for arbitrary knot sequences $\mathbf{t}$. Hence we can conclude similarly, as in the proof of the theorem, that for any $\mathbf{t}$

$$\sum_{i=1}^{n-k} \|Z_i\|^2 \leqq I_{\mathbf{t}}^2 \leqq D_k^2 \sum_{i=1}^{n-k} \|Z_i\|^2. \tag{35}$$

With the help of (35) we can show now that $I(\mathbf{z})>0$ for the data (31). Namely, we can easily check that in this case

$$\begin{aligned} Z_1 &= \sqrt{24}[t_1, t_2, t_3, t_4]\mathbf{z} \\ &= \sqrt{6}\left[\frac{(\mathbf{z}_4-\mathbf{z}_3)/h_3-(\mathbf{z}_3-\mathbf{z}_2)/h_2}{h_3+h_2} - \frac{(\mathbf{z}_3-\mathbf{z}_2)/h_2-(\mathbf{z}_2-\mathbf{z}_1)/h_1}{h_2+h_1}\right] \end{aligned}$$

where $h_i = t_{i+1}-t_i$ and $-1=t_1<t_2<t_3<t_4=1$. For the data (31) the first component is given by

$$-\sqrt{24}\left[\frac{1}{h_2(h_3+h_2)}+\frac{1}{h_2(h_2+h_1)}\right] < -\sqrt{6} < 0.$$

We therefore have $\inf_{\mathbf{t}} \|Z_1\|^2 > 0$ so that by (35) it follows that $I(z)>0$.

By the nature of our example (31) we wonder whether this phenomenon can occur in the case of one-dimensional data. Instead of investigating this question further we show here existence on the basis of the approach (32)-(34). To this end, we assume first that the data are "rough" in the sense that

$$(z_{i+1}-z_i)(z_i-z_{i-1})<0, \qquad 2 \leqq i \leqq n-1. \tag{36}$$

Visually this means the data oscillate as much as possible. In this case the divided differences in (33) have, by introducing $w_i \equiv (z_{i+1}-z_i)/(t_{i+1}-t_i)$, the form

$$[t_j, \cdots, t_{j+k}]\mathbf{z} = [t_j, \cdots, t_{j+k-1}]\mathbf{w} = \sum_{i=j}^{j+k-1} w_i \Big/ \prod_{l\neq i} (t_i - t_l).$$

Since both the w_i (by (36)) and the factors $\prod_{l\neq i}(t_i-t_l)$ alternate in sign we can write

$$|Z_j| = [(t_{j+k}-t_j)/k]^{1/2} \sum_{i=j}^{j+k-1} |w_i| \Big/ \prod_{l\neq i} |t_i - t_l|. \tag{37}$$

Now let $\mathbf{t}_N$ be, as in the proof of the theorem, a sequence of knot sequences such that $I_{\mathbf{t}_N} \to I$, $N \to \infty$. By (35) we then see that the $|Z_j|$ formed with respect to $\mathbf{t}_N$ are uniformly bounded in N. Then (37) implies that $\min_i (t^{(N)} - t_i^{(N)})$ must be uniformly bounded from below with respect to N. Hence the $\{t_i^{(N)}\}_{i=1}^n$ tend to a limiting sequence $\{t_i^*\}_{i=1}^n$ with simple knots, and existence follows as in theorem.

The case of nonrough data can be reduced to the case (36) as follows. If (36) is violated for just $i = i_0+1, \cdots, i_1$, say, then the data $z_{i_0}, \cdots, z_{i_1+1}$ are strictly monotone increasing or decreasing. We then consider a new problem ($\tilde{3}$) for the data sequence where the $z_{i_0+1}, \cdots, z_{i_1}$ are eliminated. These data then satisfy (36), and the reduced problem has a solution, which is a continuous spline function $\tilde{s}$ assuming the values z_{i_0} and z_{i_1+1} at some knots $t_{i_0}^*$ and $t_{i_1+1}^*$. Hence we can find values $t_{i_0+1}^* < \cdots < t_{i_1}^*$ in

$(t_{i_0}^*, t_{i_1+1}^*)$ such that $\tilde{s}(t_i^*) = z_i$ for $i = i_0+1, \cdots, i_1$. Since the value for I in problem $(\tilde{3})$ is a priori smaller than or equal to that of the original problem (3), we have found a solution for this latter problem also. Finally, we see that the case when (36) is violated for several strings of indices can be handled similarly.

We mention further that existence can also be proved without the additional assumption of nonasymptotic polynomial data when $k=2$. Indeed, in this case the conclusion $r \geqq 2$ required in the proof of the theorem for any distribution of t_i^*'s as in (26) is trivial.

3. Additional remarks. Further questions concerning the minimization problem (3) are the characterization and uniqueness of the solution.

The first question can be treated conveniently by variations analysis, which has been done in case $k=2$ in [3]. It can be easily extended to all k as follows. With the help of Lagrange multipliers, problem (3) is equivalent to minimizing the functional

$$\phi(\mathbf{t}, \boldsymbol{\lambda}, f) := \|\|\mathbf{f}^{(k)}\|\|^2 + \sum_{i=1}^{n} (\mathbf{f}(t_i) - \mathbf{z}_i) \cdot \boldsymbol{\lambda}_i, \qquad \boldsymbol{\lambda}_i \in \mathbb{R}^d. \tag{38}$$

Necessary conditions for this are obtained by differentiating

$$0 = \frac{\delta \phi}{\delta t_i} = \boldsymbol{\lambda}_i \cdot \mathbf{f}'(t_i), \qquad 2 \leqq i \leqq n-1, \tag{39}$$

$$0 = \frac{\delta \phi}{\delta \boldsymbol{\lambda}_i} = \mathbf{f}(t_i) - \mathbf{z}_i, \tag{40}$$

$$0 = \frac{\delta \phi}{\delta \mathbf{f}}(\mathbf{h}) = 2 \int_a^b \mathbf{f}^{(k)}(x) \cdot \mathbf{h}^{(k)}(x)\, dx + \sum_{i=1}^{n} \boldsymbol{\lambda}_i \cdot \mathbf{h}(t_i). \tag{41}$$

Here and in (38) all products are to be understood as scalar products in $\mathbb{R}^d$. Equation (41) is obtained by taking the Gateaux derivative in the sense that

$$[\phi(\mathbf{t}, \boldsymbol{\lambda}, \mathbf{f} + \delta \cdot \mathbf{h}) - \phi(\mathbf{t}, \boldsymbol{\lambda}, \mathbf{f})]/\delta \to 0, \qquad \delta \to 0.$$

Now condition (40) is just the interpolating condition in (3). Choosing the variational curves (41) as elements of $U_{k,\mathbf{t}}(\{0\})$ (cf. (2)), we show that $\mathbf{f}^{(k)}$ lies (componentwise) in the orthogonal complement of this space, and hence is a spline curve of degree $<k$.

Next we choose $\mathbf{h}_i \geqq 0$ in C^∞ with $\mathbf{h}_i(t_j) = \mathbf{e}_i$ where the $\mathbf{e}_i$ are the standard unit vectors in $\mathbb{R}^d$ and support in $(t_j - \delta, t_j + \delta)$ with $\delta \ll 1$. This yields

$$2 \int_{t_{j-1}}^{t_{j+1}} \mathbf{f}^{(k)} \mathbf{h}_i^{(k)} + \boldsymbol{\lambda}_j \cdot \mathbf{e}_i = 0.$$

Integrating by parts $k-1$ times leads to

$$2(-1)^{k-1} \int_{t_j-\delta}^{t_j+\delta} \mathbf{f}^{(2k-1)}\, d\mathbf{h}_i + \boldsymbol{\lambda}_j \cdot \mathbf{e}_i = 0.$$

Thus one more integration by parts yields

$$2(-1)^k [\mathbf{f}^{(2k-1)}]_{t_j-}^{t_j+} + \boldsymbol{\lambda}_i = 0.$$

Solving for λ_i and inserting into (39), we obtain

$$\mathbf{f}'(t_j) \cdot (\mathbf{f}^{(2k-1)}(t_{j+}) - \mathbf{f}^{(2k-1)}(t_j - 0)) = 0, \qquad 2 \leqq k \leqq n-1. \tag{42}$$

In the scalar case $d=1$ these equations imply that $\mathbf{f}$ is either **one** polynomial of degree $2k-1$ on (t_{j-1}, t_{j+1}) or else $\mathbf{f}'(t_j) = 0$. We can interpret this as an additional smoothness

property resulting from minimizing with respect to the knots. In the case $k=2$ this can be used for a uniqueness proof, which will not be produced here in view of the proof in [2].

Another approach for determining an optimal set of knots could be based on formula (17) obtained for $I_{\mathbf{t}}$ by minimizing this functional with respect to $\mathbf{t}$. However, there is the alternative formulation (34) made possible by using B-splines. On the basis of this, successful algorithms have been developed by Marin [2].

We conclude with some remarks concerning the uniqueness of a solution of (3). There is an easy counterexample showing that there will be no positive answer to this without additional assumptions. Suppose that the data $\mathbf{z}_1, \cdots, \mathbf{z}_n$ are points of a polynomial curve $\mathbf{p} \in \pi_{[k/2]}$ on the interval $[0, 1]$. Then the infimum in (3) is trivially equal to zero and attained by $\mathbf{p}(t)$ as well as by the function $\mathbf{p}(t^2) \in \pi_k$. But the latter function also interpolates the data, namely at points $t_i^* = \sqrt{t_i}$ where $\mathbf{p}(t_i) = \mathbf{z}_i$. Therefore we make the following conjecture.

CONJECTURE. If there is at most one polynomial curve in π_k interpolating the data $\{z_i\}_{i=1}^n$ at some point sequence $\mathbf{t} = \{t_i\}_{i=1}^n$ of form (1), the solution of (3) is unique (we assume (5)).

By definition this condition is necessary for uniqueness. In the special case $k=2$ of cubic spline curves, it would imply that there is always a unique solution since π_k then consists of straight lines. This is in agreement with the numerical experience gained in this case.

REFERENCES

[1] C. DE BOOR, *Splines as linear combinations of* B-*splines. A survey*, in Proc. of Conference on Approximation Theory II, G. G. Lorentz, C. K. Chui, and L. L. Schumaker, eds., Austin, TX, 1976.

[2] S. MARIN, *An approach to data parametrization in parametric cubic spline interpolation problems*, J. Approx. Theory, 41 (1984), pp. 64–86.

[3] H. J. TÖPFER, *Models for smooth curve fitting*, preprint.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

013

REPEATED INTEGRALS AND DERIVATIVES OF *K* BESSEL FUNCTIONS*

DONALD E. AMOS†

Abstract. Repeated integrals of Bessel functions $K_\nu(x)$ on $0<x<\infty$, denoted by $Ki_{\nu,n}(x)$, are considered. Series are derived in terms of exponential integrals that are direct extensions of known results for the Bickley functions $Ki_n(x)$. The basic result for $n=0$ can also be extended to $n<0$ by repeated differentiation. For ν a nonnegative integer, it is also shown that $Ki_{\nu,n}(x)$ can be represented as a finite sum of Bickley functions of which $Ki_{\nu,0}(x)=K_\nu(x)$ is a special case.

Key words. *K* Bessel functions, modified Bessel functions, Bickley functions, derivatives of Bessel functions, exponential integrals

AMS(MOS) subject classification. 33A40

1. Introduction. The Bickley functions are defined as repeated integrals of the K_0 Bessel function [1]:

$$Ki_0(x)=K_0(x), \quad Ki_n(x)=\int_x^\infty Ki_{n-1}(t)\,dt, \quad n\geq 1, \quad x>0. \tag{1.1}$$

These functions have application in heat convection problems, neutron transport, and nuclear reactor codes [4]. Reference [1, Chap. 11] gives a formula for $Ki_1(x)$ in terms of modified Struve functions, asymptotic approximations for large and small $x>0$, and citations to several tables for $Ki_n(x)$. Reference [4] summarizes the classical formulae and presents a number of highly accurate rational Chebyshev approximations designed to cover the range $n=1$ through $n=10$ and all $x>0$. The results of [3] give some basic inequalities and asymptotic expansions satisfied by the Bickley functions, as well as a computational technique to sum slowly convergent series of exponential integrals $E_k(x)$, $k\geq 0$. In particular, for $x>0$,

$$K_0(x)=\sum_{k=0}^{\infty} A_k E_{2k+1}(x), \quad Ki_n(x)=\sum_{k=0}^{\infty} A_k E_{2k+n+1}(x), \quad n\geq 0, \tag{1.2a}$$

are considered and implemented in Algorithm 609 [3] where

$$A_0=1, \quad A_k=\frac{(\frac{1}{2})_k}{k!}, \quad k\geq 1,$$
$$(a)_0\equiv 1, \quad (a)_k=a(a+1)\cdots(a+k-1), \quad k\geq 1, \tag{1.2b}$$

and $E_k(x)$ is defined in (2.1). Equation (1.2) also holds for $n<0$ since

$$Ki_{-n-1}(x)=-\frac{d}{dx}Ki_{-n}(x), \quad E_{-n-1}(x)=-\frac{d}{dx}E_{-n}(x), \quad n\geq 0.$$

These relations express the fact that $Ki_{-n}(x)$ and $E_{-n}(x)$ are, except for sign, repeated derivatives of $K_0(x)$ and $E_0(x)$, respectively. Notice that only a finite number of

* Received by the editors August 31, 1987; accepted for publication March 1, 1988. This work was supported by the U.S. Department of Energy under contract DE-AC04-76DP00789.

† Numerical Mathematics Division 1422, Sandia National Laboratories, Albuquerque, New Mexico 87185.

negative indices occurs on the exponential integrals in (1.2) when $n<0$. Functions of negative indices are conveniently generated by the following recurrences:

$$E_0(x)=e^{-x}/x, \quad E_{k-1}(x)=[e^{-x}-(k-1)E_k(x)]/x, \quad k\leqq 0, \tag{1.3}$$

$$Ki_{k-3}(x)=Ki_{k-1}(x)+[(k-1)Ki_k(x)-(k-2)Ki_{k-2}(x)]/x, \qquad k\leqq 2, \tag{1.4}$$

starting with values $Ki_2(x)$, $Ki_1(x)$, and $Ki_0(x)$ in (1.4). Stability with numerical recurrence is achieved by forward or backward recurrence away from the index $k=[x]$ where $[x]$ denotes the integer part of x [4], [6].

It is apparent from (2.3) and (3.1) that

$$Ki_{-n}(x)=(-1)^n K_0^{(n)}(x), \qquad n=0, 1, 2, \cdots. \tag{1.5}$$

That this result is consistent with (1.4) can be verified directly in the cases $k=1$ and 2, and by repeated differentiation of the differential equation

$$xK_0''(x)=-K_0'(x)+xK_0(x)$$

for other values of k.

The main results presented below extend (1.1) and (1.2) to repeated integrals of $K_\nu(x)$ defined by

$$Ki_{\nu,0}(x)=K_\nu(x), \quad Ki_{\nu,n}(x)=\int_x^\infty Ki_{\nu,n-1}(t)\,dt, \quad n\geqq 1, x>0, \nu\geqq 0. \tag{1.6}$$

Luke defines these integrals in [8] and cites a variety of formulas in [8]-[10]. We use the notation $Ki_{\nu,n}(x)$ in keeping with the notation for the Bickley functions $Ki_{0,n}(x)=Ki_n(x)$, although [3] and [8]-[10] use $K_{n,\nu}(x)$ for these functions.

In § 2, we derive a relation reciprocal to (1.2) and introduce some notation and formulae required in later sections. In § 3, we use the similarity between integral representations of $E_n(x)$ and $Ki_n(x)$ to derive series for $Ki_{\nu,n}(x)$ when ν is not an integer. The extension of these series to make ν an integer is carried out in § 4. In § 5, we derive bounds on the coefficients of the series and show asymptotic rates of convergence. While only real values of order and argument are considered in this paper, most of the results can be extended to complex orders and arguments.

2. Exponential integrals in terms of Bickley functions. We now derive a relation reciprocal to that in (1.2). Write the exponential integral $E_\nu(x)$ in the form

$$E_\nu(x)\equiv\int_1^\infty \frac{e^{-xt}}{t^\nu}\,dt=\int_0^\infty \frac{e^{-x\cosh\theta}\sinh\theta}{\cosh^\nu\theta}\,d\theta, \qquad \nu>0, \tag{2.1}$$

and substitute

$$\sinh\theta=\cosh\theta\sqrt{1-\frac{1}{\cosh^2\theta}}=\sum_{k=0}^\infty B_k\frac{1}{\cosh^{2k-1}\theta}, \qquad B_0=1, \quad B_k=\frac{(-1/2)_k}{k!}, \quad k\geqq 1 \tag{2.2}$$

into (2.1). Then, using an integral representation for $Ki_\nu(x)$ [1, p. 483],

$$Ki_\nu(x)=\int_0^\infty \frac{e^{-x\cosh\theta}}{\cosh^\nu\theta}\,d\theta, \qquad x\geqq 0, \nu\geqq 0, x+\nu>0, \tag{2.3}$$

we have exponential integrals in terms of Bickley functions

$$E_\nu(x)=\sum_{k=0}^\infty B_k Ki_{2k+\nu-1}(x). \tag{2.4}$$

We can also see from the integral forms (2.1) and (2.3) that (1.2) and (2.4) are valid if we do repeated differentiation (ν a negative integer) on the case $\nu = 0$.

3. Generalizations of (1.2). In this section we exploit (2.1) and (2.3) to derive series representations for $K_\nu(x)$ and $Ki_{\nu,n}(x)$. We start with a well-known integral representation for $K_\nu(x)$ [1, p. 376]:

$$K_\nu(z) = \int_0^\infty e^{-z\cosh\theta} \cosh \nu\theta \, d\theta, \qquad z > 0, \tag{3.1}$$

and write $K_\nu(z)$ in the form $K_\nu(z) = U_\nu(z) + L_\nu(z)$ where

$$U_\nu(z) = \frac{1}{2}\int_0^\infty e^{-z\cosh\theta+\nu\theta}\, d\theta, \qquad L_\nu(z) = \frac{1}{2}\int_0^\infty e^{-z\cosh\theta-\nu\theta}\, d\theta. \tag{3.2}$$

In order to carry out the derivation, we need the following relations [5, p. 101]:

$$e^{-\nu\theta} = \frac{\sinh\theta}{2^\nu \cosh^{\nu+1}\theta} F\left[1+\frac{\nu}{2}, \frac{1+\nu}{2}; 1+\nu; \frac{1}{\cosh^2\theta}\right], \qquad 0 < \theta < \infty, \tag{3.3}$$

$$e^{-\nu\theta} = \frac{1}{2^\nu \cosh^\nu\theta} F\left[\frac{\nu}{2}, \frac{1+\nu}{2}; 1+\nu; \frac{1}{\cosh^2\theta}\right], \qquad 0 \leqq \theta < \infty, \tag{3.4}$$

for the exponential terms in $U_\nu(z)$ and $L_\nu(z)$ where F is the Gauss hypergeometric function. Equations (3.3) and (3.4) are related by the Kummer transformation [1, eq. (15.3.3)]. These relations can be derived from equations (15.1.1), (15.3.23), and (15.3.3) of [1]. Notice that the parameterization of $F(a, b; c; z)$ in these cases has a relation $a - b + \frac{1}{2} = 0$ that makes [1, eq. (15.3.23)] reduce to

$$F\left[\frac{\nu}{2}, \frac{1+\nu}{2}; 1+\nu; z\right] = \left[\frac{2}{1+\sqrt{1-z}}\right]^\nu, \qquad |z| < 1. \tag{3.5}$$

The Kummer relation (5.2) [1, eq. (15.3.3)] applied to this result gives

$$F\left[1+\frac{\nu}{2}, \frac{1+\nu}{2}; 1+\nu; z\right] = \frac{1}{\sqrt{1-z}}\left[\frac{2}{1+\sqrt{1-z}}\right]^\nu, \qquad |z| < 1, \tag{3.6}$$

and with $z = 1/\cosh^2\theta$ we get (3.3) and (3.4). Notice that the series form of (3.3) is correct as $\theta \to \infty$, but does not converge at $\theta = 0$, making (3.3) indeterminant there. The proper form near the origin is given by the power series form of (3.4) (see § 5 for rates of convergence).

To derive series expansions for (3.1) we expand the right-hand side of (3.3) in a power series using the definition

$$F(a, b; c; z) = \sum_{k=0}^\infty \frac{(a)_k (b)_k}{(c)_k k!} z^k, \qquad |z| < 1, \tag{3.7}$$

and substitute the result in the second form of (3.2). With the help of (2.1) we have

$$L_\nu(z) = \sum_{k=0}^\infty A_k(\nu) \int_0^\infty \frac{e^{-z\cosh\theta}\sinh\theta}{\cosh^{2k+\nu+1}\theta}\, d\theta = \sum_{k=0}^\infty A_k(\nu) E_{2k+\nu+1}(z). \tag{3.8}$$

Similarly, if we use (3.4) in place of (3.3), we obtain, using (2.3),

$$L_\nu(z) = \sum_{k=0}^\infty B_k(\nu) Ki_{2k+\nu}(z) \tag{3.9}$$

where $A_k(\nu)$ and $B_k(\nu)$ are defined by (3.3) and (3.4):

$$A_0(\nu)=2^{-\nu}, \qquad B_0(\nu)=2^{-\nu},$$

(3.10)

$$A_k(\nu)=\left(1+\frac{\nu}{2}\right)_k\left(\frac{1+\nu}{2}\right)_k\Big/2^\nu k!(1+\nu)_k, \quad B_k(\nu)=\left(\frac{\nu}{2}\right)_k\left(\frac{1+\nu}{2}\right)_k\Big/2^\nu k!(1+\nu)_k, \quad k\geqq 1.$$

Notice that $A_k(0)$ is the A_k of (1.2), but $B_k(0)$ is not the B_k of (2.2). To compute the dominant or upper piece of (3.2) we need $e^{\nu\theta}$. Formally, we write (3.3) and (3.4) with ν replaced by $-\nu$ and consider noninteger values for ν. Then,

$$e^{\nu\theta}=\cosh^{\nu-1}\theta\sinh\theta\sum_{k=0}^{\infty}\frac{C_k(\nu)}{\cosh^{2k}\theta}, \qquad e^{\nu\theta}=\cosh^{\nu}\theta\sum_{k=0}^{\infty}\frac{D_k(\nu)}{\cosh^{2k}\theta}, \tag{3.11}$$

where

(3.12)

$$C_k(\nu)=2^\nu\left(1-\frac{\nu}{2}\right)_k\left(\frac{1-\nu}{2}\right)_k\Big/k!(1-\nu)_k, \qquad D_k(\nu)=2^\nu\left(\frac{-\nu}{2}\right)_k\left(\frac{1-\nu}{2}\right)_k\Big/k!(1-\nu)_k.$$

Here ν is considered nonintegral because the terms of (3.12) become indeterminant for $k\geqq\nu$ when ν is an integer. We shall resolve this indeterminacy in § 4. The reciprocals can be taken because the closed forms obtained from (3.5) and (3.6),

$$\sum_{k=0}^{\infty}B_k(\nu)z^k=\left[\frac{1}{1+\sqrt{1-z}}\right]^\nu, \qquad \sum_{k=0}^{\infty}A_k(\nu)z^k=\frac{1}{\sqrt{1-z}}\left[\frac{1}{1+\sqrt{1-z}}\right]^\nu, \tag{3.13}$$

show a radius of convergence $|z|<1$ and neither has any zeros in $|z|<1$. Thus, applying (3.11) to the dominant or upper part of (3.1) yields

$$U_\nu(z)=\sum_{k=0}^{\infty}C_k(\nu)E_{2k-\nu+1}(z) \quad\text{and}\quad U_\nu(z)=\sum_{k=0}^{\infty}D_k(\nu)Ki_{2k-\nu}(z), \tag{3.14}$$

after using the integral representations of $E_\nu(z)$ and $Ki_\nu(z)$ from § 2.

To summarize the derivation, (3.1) reduces to

$$K_\nu(z)=\begin{cases}\dfrac{1}{2}\displaystyle\sum_{k=0}^{\infty}C_k(\nu)E_{2k-\nu+1}(z)+\frac{1}{2}\sum_{k=0}^{\infty}A_k(\nu)E_{2k+\nu+1}(z),\\[2ex] \dfrac{1}{2}\displaystyle\sum_{k=0}^{\infty}D_k(\nu)Ki_{2k-\nu}(z)+\frac{1}{2}\sum_{k=0}^{\infty}B_k(\nu)Ki_{2k+\nu}(z).\end{cases} \tag{3.15}$$

Since $E_\nu(z)$ satisfies an integral recurrence such as (1.1), we get, with repeated integrations ($n\geqq 0$) or repeated differentiations ($n<0$) on (3.15),

$$Ki_{\nu,n}(z)=\begin{cases}\dfrac{1}{2}\displaystyle\sum_{k=0}^{\infty}C_k(\nu)E_{2k-\nu+n+1}(z)+\frac{1}{2}\sum_{k=0}^{\infty}A_k(\nu)E_{2k+\nu+n+1}(z),\\[2ex] \dfrac{1}{2}\displaystyle\sum_{k=0}^{\infty}D_k(\nu)Ki_{2k-\nu+n}(z)+\frac{1}{2}\sum_{k=0}^{\infty}B_k(\nu)Ki_{2k+\nu+n}(z).\end{cases} \tag{3.16}$$

It is apparent from repeated integrations of (3.1) that $Ki_{\nu,n}(z)$ has an integral representation

$$Ki_{\nu,n}(z)=\int_0^\infty\frac{e^{-z\cosh\theta}\cosh\nu\theta}{\cosh^n\theta}\,d\theta, \qquad \operatorname{Re}z>0, \quad \nu\geqq 0, \tag{3.17}$$

which generalizes (2.3). Repeated differentiation of this relation for $n=0$ produces

$$Ki_{\nu,-n-1}(z)=-\frac{d}{dz}Ki_{\nu,-n}(z)=(-1)^{n+1}\frac{d^{n+1}}{dz^{n+1}}K_\nu(z). \tag{3.18}$$

4. Identification of $C_k(\nu)$ and $D_k(\nu)$ for $\nu=m$ a nonnegative integer. We noted in § 3 that ν had to be nonintegral in order to make sense of (3.15) because the coefficients of the hypergeometric relation become indeterminant for $k\geqq\nu$ when ν is a nonnegative integer. Thus, for ν not an integer, we write $C_k(\nu)$ in terms of gamma functions

$$\begin{aligned}C_k(\nu)=2^\nu\Gamma\left(1-\frac{\nu}{2}+k\right)\Gamma\left(\frac{1-\nu}{2}+k\right)\Big/\Gamma(k+1)\Gamma(1-\nu+k)\\ \cdot\Gamma(1-\nu)\Big/\Gamma\left(1-\frac{\nu}{2}\right)\Gamma\left(\frac{1-\nu}{2}\right),\end{aligned} \tag{4.1}$$

with similar manipulations for $D_k(\nu)$. The duplication relation for $\Gamma(2z)$ [1, p. 256] applied to $\Gamma(1-\nu)$ resolves the indeterminacy. At this point ν can be considered an integer m and the substitution $k=m+s$, $s=0, 1, \cdots$ gives

$$C_{m+s}(m)=\frac{1}{\sqrt{\pi}}\left(\Gamma\left(1+\frac{m}{2}+s\right)\Gamma\left(\frac{1+m}{2}+s\right)\Big/\Gamma(1+m+s)\Gamma(1+s)\right)=A_s(m),\qquad s\geqq 0. \tag{4.2}$$

Similar manipulations on $D_k(\nu)$ for $k\geqq\nu$ give $D_{m+s}(m)=-B_s(m)$, $s\geqq 0$. To summarize, we have

$$C_s(m)=\begin{cases}A_s(-m), & s<m,\\ A_{s-m}(m), & s\geqq m,\end{cases}\qquad D_s(m)=\begin{cases}B_s(-m), & s<m,\\ -B_{s-m}(m), & s\geqq m.\end{cases} \tag{4.3}$$

Thus, (3.16) reduces to

$$Ki_{m,n}(z)=\begin{cases}\frac{1}{2}\sum_{k=0}^{m-1}A_k(-m)E_{2k-m+n+1}(z)+\sum_{k=0}^{\infty}A_k(m)E_{2k+m+n+1}(z), & \text{(4.4a)}\\ \frac{1}{2}\sum_{k=0}^{m-1}B_k(-m)Ki_{2k-m+n}(z), & \text{(4.4b)}\end{cases}\qquad m\geqq 1.$$

If we specialize m or n, we get the following special cases:

$$\begin{aligned}&Ki_{1,n}(z)=Ki_{n-1}(z),\qquad Ki_{1,1}(z)=Ki_0(z)=K_0(z),\\ &Ki_{1,0}(z)=Ki_{-1}(z)=K_1(z),\\ &Ki_{m,0}(z)=K_m(z)=\frac{1}{2}\sum_{k=0}^{m-1}B_k(-m)Ki_{2k-m}(z).\end{aligned} \tag{4.5}$$

Equation (4.4b) shows that $Ki_{\nu,n}(z)$ can be computed as a finite sum of Bickley functions when ν is a nonnegative integer. Reference [3] provides a Bickley function code for nonnegative indices while (1.4) extends these results to negative indices.

In (4.4a), the upper limit $m-1$ is actually $(m/2)-1$ if m is even and $(m-1)/2$ if m is odd because the succeeding coefficients are zero. The corresponding upper limit in (4.4b) is $(m/2)$ if m is even and $(m-1)/2$ if m is odd. This can be seen from (3.10) by replacing ν by $-m$. There is an index k such that one of the terms (as well as

succeeding terms) will be zero. For this index, $A_k(-m)$ and $B_k(-m)$ from (3.11) terminate with one of the following factors:

$$\begin{array}{lll} & m \text{ even} & m \text{ odd} \\ A_k(-m): & 1-\frac{m}{2}+k-1=0 \quad \text{or} & \frac{1-m}{2}+k-1=0, \\ B_k(-m): & -\frac{m}{2}+k-1=0 \quad \text{or} & \frac{1-m}{2}+k-1=0. \end{array}$$

Since these values of k yield the first zero term, the index of the last nonzero term is one less. This shows that the sum for $K_m(z)$ in (4.5) and its derivatives ($n<0$) in (4.4b) contain only Bickley functions of nonpositive indices which, according to (1.5), means only derivatives of $K_0(z)$ are involved.

5. Bounds and asymptotic estimates for $A_k(\nu)$ and $B_k(\nu)$, $\nu\geqq 0$, $k\geqq 0$. The relation

$$B_n(\nu)=\sum_{k=0}^{n} A_k(\nu)B_{n-k}, \qquad n\geqq 0, \quad \nu\geqq 0 \tag{5.1}$$

is derived from (3.7) by substituting the power series (3.7) into the Kummer relation

$$F\left(\frac{\nu}{2}, \frac{1+\nu}{2}; 1+\nu; z\right)=(1-z)^{1/2}F\left(1+\frac{\nu}{2}, \frac{1+\nu}{2}; 1+\nu; z\right), \qquad \nu\geqq 0, \tag{5.2}$$

and equating like powers of z. Multiplication of (5.2) by $(1-z)^{-1}$ and equating like powers of z again gives

$$\sum_{k=0}^{n} B_k(\nu)=\sum_{k=0}^{n} A_k(\nu)A_{n-k}(0), \qquad n\geqq 0. \tag{5.3}$$

On the other hand, multiplication of (5.2) by $(1-z)^{-1/2}$ gives

$$\sum_{k=0}^{n} B_k(\nu)A_{n-k}(0)=A_n(\nu), \qquad n\geqq 0, \tag{5.4}$$

and with $A_k(0)\leqq 1$, $k\geqq 0$, we get

$$A_k(\nu)\leqq \sum_{k=0}^{n} B_k(\nu), \qquad n\geqq 0.$$

Now, (2.2) shows that $B_0=1$ and $B_k<0$ for $k\geqq 1$. Therefore, (5.1) shows that

$$B_n(\nu)\leqq A_n(\nu), \qquad n\geqq 0. \tag{5.5}$$

Equation (3.13) for $z=1$ gives

$$\sum_{k=0}^{n} B_k(\nu)<\sum_{k=0}^{\infty} B_k(\nu)=1, \qquad \nu\geqq 0, \quad n\geqq 0, \tag{5.6}$$

and combining (5.3)-(5.6) we get

$$B_n(\nu)\leqq A_n(\nu)\leqq \sum_{k=0}^{n} B_k(\nu)=\sum_{k=0}^{n} A_k(\nu)A_{n-k}(0)<1, \qquad \nu\geqq 0, \quad n\geqq 0. \tag{5.7}$$

In order to estimate the rate of convergence of the series in (4.4) we need asymptotic estimates of the coefficients. The asymptotic estimate [1, p. 257]

$$\frac{\Gamma(z+a)}{\Gamma(z+b)}\sim z^{a-b}, \qquad z\to\infty, \tag{5.8}$$

applied to $A_k(\nu)$ and $B_k(\nu)$ gives

$$(5.9)\qquad A_k(\nu)\sim\frac{1}{\sqrt{\pi}}k^{-1/2}=O(k^{-1/2}),\quad B_k(\nu)\sim\frac{1}{\sqrt{\pi}}k^{-3/2}=O(k^{-3/2})\quad\text{as } k\to\infty.$$

Equation (5.9) and the estimates [1, p. 229]

$$\frac{e^{-z}}{z+k}<E_k(z)\leqq\frac{e^{-z}}{z+k-1},\quad k=1,2,\cdots,\quad z\geqq 0,$$

on the terms of (4.4a) give an order relation $O(k^{-3/2})$ as $k\to\infty$. Consequently, the series after summing N terms has a truncation error $O(N^{-1/2})$ as $N\to\infty$. Reference [3] deals with the numerical problems of summing slowly convergent series of exponential integrals. These results can be applied directly to that part of (4.4a) where $2k+m+n+1>0$ provided that the truncation error bound in [3] uses $A_n(m)<1$ from (5.7). The corresponding changes for Algorithm 609 [3] would be trivial.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions*, Applied Mathematics Series 55, National Bureau of Standards, Washington, DC, December 1965.

[2] D. E. AMOS, *Computation of exponential integrals*, ACM Trans. Math. Software, 6 (1980), pp. 365-377; Algorithm 556, Exponential integrals, pp. 420-428.

[3] ———, *Uniform asymptotic expansions for exponential integrals $E_n(x)$ and Bickley Functions $Ki_n(x)$*, ACM Trans. Math. Software, 9 (1983), pp. 467-479; Algorithm 609, A portable FORTRAN subroutine for the Bickley functions $Ki_n(x)$, pp. 480-493.

[4] J. M. BLAIR, C. A. EDWARDS, AND J. H. JOHNSON, *Rational Chebyshev approximations for the Bickley functions $Ki_n(x)$*, Math. Comp., 32 (1978), pp. 876-886.

[5] A. ERDELYI, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[6] W. GAUTSCHI, *Exponential integrals*, Commun. ACM, 16 (1973), pp. 761-763.

[7] ———, *Exponential integral $E_n(x)$ for large values of n*, J. Res. Nat. Bur. Standards, 62 (1959), pp. 123-125.

[8] Y. L. LUKE, *Integrals of Bessel Functions*, McGraw-Hill, New York, 1962.

[9] ———, *Mathematical Functions and Their Approximations*, Academic Press, New York, 1975.

[10] ———, *The Special Functions and Their Approximations*, Vol. II, Academic Press, New York, 1969.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

014

A POSITIVE TRIGONOMETRIC SUM*

JOAQUIN BUSTOZ† AND MOURAD E. H. ISMAIL‡

Abstract. The positivity of

$$\sum_{k=0}^{n} \frac{(\lambda)_k(\lambda)_{n-k}}{k!(n-k)!} \frac{\sin(k+1)\theta}{k+1}$$

on $[\pi/3, \pi)$ when $3<\lambda\leqq 4$ is established.

Key words. Fejer's inequality, ultraspherical polynomials

AMS(MOS) subject classification. 33A45, 42A05

1. Introduction. Let $(\lambda)_0=1$ and $(\lambda)_n=\lambda(\lambda+1)\cdots(\lambda+n-1)$, $n=1,2,\cdots$. The trigonometric inequality

$$\sum_{k=0}^{n} \frac{(\lambda)_k}{k!} \frac{(\lambda)_{n-k}}{(n-k)!} \frac{\sin(k+1)\theta}{k+1}>0, \qquad 0<\theta<\pi, \quad n=0,1,2,\cdots, \tag{1.1}$$

is known to hold for $0\leqq\lambda\leqq 3$ [1]-[3]. When $\lambda>3$, it is known that (1.1) fails for infinitely many values of n [1]. In this note we will prove that if $3<\lambda\leqq 4$, then (1.1) holds in $\pi/3\leqq\theta<\pi$ and that when $\lambda=4$ the $\pi/3$ is best possible. When $\lambda=1$, then (1.1) reduces to Fejer's classic inequality

$$\sum_{k=0}^{n} \frac{\sin(k+1)\theta}{k+1}>0, \qquad 0<\theta<\pi.$$

Let $P_n^{\lambda}(x)$ denote the ultraspherical polynomial defined by

$$(1-2xz+z^2)^{-\lambda}=\sum_{n=0}^{\infty} P_n^{\lambda}(x)z^n,$$

and set

$$T_n(\theta,\lambda)=\sum_{k=0}^{\infty} \frac{(\lambda)_k}{k!} \frac{(\lambda)_{n-k}}{(n-k)!} \frac{\sin(k+1)\theta}{k+1}.$$

Write $x=\cos\theta/2$ and set

$$\Delta_n(x,\lambda)=\sin[(n+1)\theta/2]P_{n+1}^{\lambda}(x)-\sin[(n+2)\theta/2]P_n^{\lambda}(x).$$

It has been proved in [3] that

$$T_n(\theta,\lambda)=\frac{\Delta_n(x,\lambda)}{2(\lambda-1)x}, \qquad 0<x<1,$$

and, consequently, if $\lambda>1$ positivity of $T_n(\theta,\lambda)$ follows from positivity of $\Delta_n(x,\lambda)$. In [3] it was also proved that the generating function for $P_n^{\lambda}(x)$ implies (recall

* Received by the editors November 18, 1985; accepted for publication (in revised form) March 7, 1988.
† Department of Mathematics, Arizona State University, Tempe, Arizona 85287.
‡ Department of Mathematics, University of South Florida, Tampa, Florida 33620.

$x = \cos \theta/2)$

$$\sum_{n=0}^{\infty} \Delta_n(x, \lambda) t^{n+1} = (1-t)^{-\lambda} \operatorname{Im} (1 - te^{i\theta})^{-\lambda+1}. \tag{1.2}$$

By applying Darboux's method [4] to (1.2), we find that

$$\frac{n!\Delta_{n-1}(x, \lambda)}{(\lambda)_n} = \left(2 \sin\left(\frac{\theta}{2}\right)\right)^{(1-\lambda)} \sin\left[(\lambda - 1)\left(\frac{\pi - \theta}{2}\right)\right] + O\left(\frac{1}{n}\right), \qquad 0 < \theta < \pi. \tag{1.3}$$

It follows from (1.3) that if $\lambda > 3$ and $(\lambda - 3/\lambda - 1)\pi \leqq \theta \leqq \pi$, then $T_n(\theta, \lambda)$ is positive for large n. We conjecture that $T_n(\theta, \lambda) > 0$ for $n = 0, 1, 2, \cdots$ if $\lambda > 3$, $(\lambda - 3/\lambda - 1)\pi \leqq \theta < \pi$. The lower limit $(\lambda - 3/\lambda - 1)\pi$ is clearly best possible for large n because of (1.3) (in particular at $\lambda = 4$, $\pi/3$ is best possible).

2. The case $\lambda = 4$. In this section we will prove the following theorem.

THEOREM 1.

$$\sum_{k=0}^{n} \frac{(4)_k}{k!} \frac{(4)_{n-k}}{(n-k)!} \frac{\sin(k+1)\theta}{k+1} > 0, \qquad \frac{\pi}{3} \leqq \theta < \pi, \quad n = 0, 1, 2, \cdots.$$

Proof. Set $\lambda = 4$ in (1.2). Then by partial fractions

$$\begin{aligned} \frac{1}{(1-t)^4(1-te^{i\theta})^3} &= \frac{A_1}{1-t} + \frac{A_2}{(1-t)^2} + \frac{A_3}{(1-t)^3} + \frac{A_4}{(1-t)^4} + \frac{B_1}{1-te^{i\theta}} \\ &\quad + \frac{B_2}{(1-te^{i\theta})^2} + \frac{B_3}{(1-te^{i\theta})^3}. \end{aligned} \tag{2.1}$$

Expand $(1 - te^{i\theta})^{-3}$ in a Taylor series about $t = 1$ to get

$$(1-te^{i\theta})^{-3} = (1-e^{i\theta})^{-3} \sum_{n=0}^{\infty} \frac{(3)_n}{n!} \left(\frac{e^{i\theta}}{e^{i\theta}-1}\right)^n (1-t)^n.$$

Hence, the coefficients A_i in (2.1) are given by

$$A_1 = -10\, e^{3i\theta}(e^{i\theta}-1)^{-6}, \qquad A_2 = -6\, e^{2i\theta}(e^{i\theta}-1)^{-5},$$
$$A_3 = -3\, e^{i\theta}(e^{i\theta}-1)^{-4}, \qquad A_4 = -(e^{i\theta}-1)^{-3}.$$

Expand $(1-t)^{-4}$ about $t = e^{-i\theta}$ to get

$$(1-t)^{-4}(1-te^{i\theta})^{-3} = e^{4i\theta} \sum_{n=0}^{\infty} \frac{(4)_n}{n!} (1-e^{i\theta})^{-n-4}(1-e^{i\theta}t)^{n-3}.$$

Hence, the coefficients B_i in (2.1) are given by

$$B_1 = 10\, e^{4i\theta}(1-e^{i\theta})^{-6}, \quad B_2 = 4\, e^{4i\theta}(1-e^{i\theta})^{-5}, \quad B_3 = e^{4i\theta}(1-e^{i\theta})^{-4}.$$

From (1.2) and (2.1), we have

$$\Delta_{n-1}(x, 4) = \sum_{k=1}^{4} \frac{(k)_n}{n!} \operatorname{Im} A_k + \sum_{k=1}^{3} \frac{(k)_n}{n!} \operatorname{Im} B_k\, e^{in\theta}. \tag{2.2}$$

Write $\beta = 2 \sin \theta/2$ for notational convenience. Then the imaginary parts in (2.2) are given by

$$\operatorname{Im} A_1 = 0, \qquad \operatorname{Im} A_2 = 6\beta^{-5} \cos\frac{\theta}{2},$$

$$\operatorname{Im} A_3 = 3\beta^{-4} \sin\theta, \qquad \operatorname{Im} A_4 = -\beta^{-3} \cos\frac{3\theta}{2},$$

$$\operatorname{Im} B_1\, e^{in\theta} = -10\beta^{-6} \sin(n+1)\theta, \qquad \operatorname{Im} B_2\, e^{in\theta} = 4\beta^{-5} \cos\left(n + \frac{3}{2}\right)\theta,$$

$$\operatorname{Im} B_3\, e^{in\theta} = \beta^{-4} \sin(n+2)\theta.$$

Replace these imaginary parts in (2.2) and do some minor simplification to get

$$
\begin{aligned}
96\left(\sin\frac{\theta}{2}\right)^6\Delta_{n-1}(x,4) &= -2(n+3)(n+2)(n+1)\left(\sin\frac{\theta}{2}\right)^3\cos\frac{3\theta}{2}\\
&\quad +9(n+2)(n+1)\left(\sin\frac{\theta}{2}\right)^2\sin\theta+9(n+1)\sin\theta\\
&\quad -15\sin(n+1)\theta+12(n+1)\sin\frac{\theta}{2}\cos\left(n+\frac{3}{2}\right)\theta\\
&\quad +3(n+2)(n+1)\left(\sin\frac{\theta}{2}\right)^2\sin(n+2)\theta.
\end{aligned}
\tag{2.3}
$$

The proof now consists of showing that the trigonometric polynomial on the right-hand side of (2.3) is positive in $[\pi/3, \pi)$ for $n=1, 2, 3, \cdots$. The interval $[\pi/3, \pi)$ will be broken into the three subintervals: $[\pi/3, \pi/2]$, $(\pi/2, 2\pi/3]$, $(2\pi/3, \pi)$. For notational convenience, write $D_n = 96(\sin\theta/2)^6\Delta_{n-1}(x,4)$ and denote the six terms on the right-hand side of (2.3) by $a_1, a_2, a_3, a_4, a_5, a_6$ in the order in which they appear. Thus, (2.3) is rewritten as

$$D_n = a_1+a_2+a_3+a_4+a_5+a_6.$$

First consider the interval $[\pi/3, \pi/2]$. In this interval $a_1 \geqq 0$, and thus $D_n \geqq a_2+a_3+a_4+a_5+a_6$. Clearly $|a_4| \leqq 15$, $|a_5| \leqq 12(n+1)\sin\theta/2$, and $|a_6| \leqq 3(n+2)(n+1)(\sin\theta/2^2)$. Hence,

$$D_n \geqq 3(n+1)(n+2)(3\sin\theta-1)\sin^2\frac{\theta}{2}+6(n+1)\left(3\cos\frac{\theta}{2}-2\right)\sin\frac{\theta}{2}-15. \tag{2.4}$$

For $\theta \in [\pi/3, \pi/2]$, $3(3\sin\theta-1)\sin^2\theta/2 > 1.19$ and $6(3\cos\theta/2-2)\sin\theta/2 \geqq 9-6\sqrt{2} > 0.51$. Replacing these lower bounds in (2.4), we obtain

$$D_n \geqq (1.19)(n+2)(n+1)+(0.51)(n+1)-15 > 0, \qquad n=2, 3, \cdots.$$

The case $n=1$ is trivial.

Next consider the interval $\pi/2 < \theta \leqq 2\pi/3$. In this interval, the first two terms on the right-hand side of (2.3), a_1 and a_2, are increasing. Hence, $a_1+a_2 \geqq \frac{1}{2}(n+3)(n+2)(n+1)+9/2(n+2)(n+1)$. Also, $\sin\theta \geqq \sqrt{3}/2$ in this interval, and hence $a_3 \geqq 9\sqrt{3}/2\,(n+1)$. Clearly, $a_4 \geqq -15$. Since $\sin\theta/2 \leqq \sqrt{3}/2$, we have $a_5 \geqq -6\sqrt{3}(n+1)$ and $a_6 \geqq -9/4(n+2)(n+1)$. Putting these estimates together, we have

$$4D_n \geqq 2(n+3)(n+2)(n+1)+9(n+2)(n+1)-6\sqrt{3}(n+1)-15 > 0, \qquad n=1, 2, \cdots.$$

The final subinterval $(2\pi/3, \pi)$ requires a preliminary lemma.

LEMMA 1. *If* $2\pi/3 \leqq \theta \leqq \pi$, *then*

$$\frac{\sin(n+2)\theta}{n+2} \geqq \frac{\sin 2\theta}{2} \text{ and } \frac{\cos(n+3/2)\theta}{n+3/2} \geqq \frac{\cos 3/2\theta}{3/2}$$

for $n=0, 1, 2, \cdots$.

Proof. Set $f(\theta) = \sin(n+2)\theta/(n+2)-\sin 2\theta/2$. Since $\sin(n+2)\,2\pi/3$ assumes the values $0, \sqrt{3}/2, -\sqrt{3}/2$ and since $\sin 4\pi/3 = -\sqrt{3}/2$, we see that $f(2\pi/3) \geqq 0$. Also $f(\pi)=0$. Now $f'(\theta) = \cos(n+2)\theta-\cos 2\theta = -2\sin(n+4)\theta/2\sin n\theta/2$, so that $f'(\theta) = 0$ when $\sin(n+4)\theta/2=0$ and when $\sin n\theta/2=0$. That is, at the points $\theta_k = 2k\pi/(n+4)$ and $\theta_k' = 2k\pi/n$ where $\theta_k, \theta_k' \varepsilon (2\pi/3, \pi)$. We have $f(\theta_k) = -(n+4)/2(n+2) \cdot \sin 4k\pi/(n+4) > 0$ because $4k\pi/(n+4)$ must be a third or fourth quadrant angle since $2k\pi/(n+4) \in (2\pi/3, \pi)$. Similarly, $f(\theta_k') = -n/2(n+2)\sin 4k\pi/n > 0$ because $2k\pi/n \in (2\pi/3, \pi)$. This proves the first inequality.

To prove the second inequality, set $g(\theta)=\cos(n+3/2)\theta/(n+3/2)-(\cos 3\theta/2)(3/2)$. Clearly, $g(2\pi/3)\geqq\frac{2}{3}-1/(n+3/2)\geqq 0$ for $n=1, 2, \cdots$, also $g(\pi)=0$. Now $g'(\theta)=-2\cos(n+3)\theta/2\sin n\theta/2$ and so $g'(\theta)=0$ at the points $\theta_k=(2k+1)\pi/(n+3)$ and $\theta'_k=4k\pi/n$. We consider only those values of k for which θ_k, $\theta'_k\in(2\pi/3, \pi)$. We have $f(\theta_k)=-(4n+3)(6n+9)\cos(6k+3)(2n+6)\pi$. Then $f(\theta_k)>0$ since $(6k+3)(2n+6)\pi$ is a third quadrant angle. Similarly, $f(\theta'_k)=-4n/(6n+9)\cos 6k\pi/n$ and $f(\theta'_k)>0$ since $6k\pi/n$ is a third quadrant angle. This proves the lemma.

Now we turn to the interval $(2\pi/3, \pi)$. First, by the identity $\sin\theta/2\cos 3\theta/2=\sin\theta(\cos\theta-\frac{1}{2})$ we rewrite a_1 as $a_1=-2(n+3)(n+2)(n+1)(\cos\theta-\frac{1}{2})\sin\theta\sin^2\theta/2$. Then dividing through (2.3) by $\sin\theta$ we have

$$\begin{aligned}\frac{D_n}{\sin\theta}=&-2(n+3)(n+2)(n+1)\left(\cos\theta-\frac{1}{2}\right)\sin^2\frac{\theta}{2}+9(n+2)(n+1)\sin^2\frac{\theta}{2}+9(n+1)\\ &-15\frac{\sin(n+1)\theta}{\sin\theta}+12(n+1)\left(n+\frac{3}{2}\right)\frac{\sin\theta/2}{\sin\theta}\frac{\cos(n+3/2)\theta}{n+3/2}\\ &+3(n+2)^2(n+1)\frac{\sin^2\theta/2}{\sin\theta}\frac{\sin(n+2)\theta}{n+2}.\end{aligned}\tag{2.5}$$

Next, replacing the last two terms in (2.5) in accordance with the lemma and using $|\sin(n+1)\theta/\sin\theta|\leqq n+1$ in the fourth term, we have

$$\begin{aligned}\frac{D_n}{\sin\theta}\geqq&(n+3)(n+2)(n+1)(1-2\cos\theta)\sin^2\frac{\theta}{2}+9(n+2)(n+1)\sin^2\frac{\theta}{2}\\ &+9(n+1)-15(n+1)+8(n+1)\left(n+\frac{3}{2}\right)\frac{\sin\theta/2\cos 3\theta/2}{\sin\theta}\\ &+3(n+2)^2(n+1)\sin^2\frac{\theta}{2}\cos\theta.\end{aligned}\tag{2.6}$$

In the fifth term in (2.6) again write $\sin\theta/2\cos 3\theta/2=\sin\theta(\cos\theta-\frac{1}{2})$ and divide through by $n+1$ to get

$$\begin{aligned}\frac{D_n}{(n+1)\sin\theta}\geqq&(n+3)(n+2)(1-2\cos\theta)\sin^2\frac{\theta}{2}+9(n+2)\sin^2\frac{\theta}{2}\\ &+8\left(n+\frac{3}{2}\right)\left(\cos\theta-\frac{1}{2}\right)+3(n+2)^2\sin^2\frac{\theta}{2}\cos\theta-6.\end{aligned}\tag{2.7}$$

Multiplying out the factors involving n in (2.7) and collecting terms gives

$$\frac{D_n}{(n+1)\sin\theta}\geqq\left[(1+\cos\theta)\sin^2\frac{\theta}{2}\right]n^2+\left[6\sin^2\frac{\theta}{2}+4\cos\theta+2\sin^2\frac{\theta}{2}\cos\theta\right]n.\tag{2.8}$$

The trigonometric polynomial in the second term in (2.8) can be rewritten as $(3-\cos\theta)(1+\cos\theta)$.

Replacing this in (2.8) and dividing by $n(1+\cos\theta)$ gives

$$\frac{D_n}{n(n+1)(1+\cos\theta)\sin\theta}\geqq n\sin^2\frac{\theta}{2}+3-\cos\theta>0.\tag{2.9}$$

This completes the proof of Theorem 1.

3. The case $3<\lambda<4$. In this section we will prove the following positivity result.

THEOREM 2. *$T_n(\theta, \lambda)>0$ for $\pi/3\leqq\theta<\pi$, $3<\lambda<4$, $n=0, 1, 2, \cdots$.*

Proof. The inequality proved in Theorem 1 is equivalent to

$$P_n^1(x)P_{n+1}^4(x)-P_{n+1}^1(x)P_n^4(x)>0, \qquad 0<x\leqq\frac{\sqrt{3}}{2}, \quad n=0, 1, 2, \cdots \tag{3.1}$$

and the inequality in Theorem 2 is equivalent to

$$P_n^1(x)P_{n+1}^\lambda(x)-P_{n+1}^1(x)P_n^\lambda(x)>0, \qquad 0<x\leqq\frac{\sqrt{3}}{2}, \quad 3<\lambda<4, \quad n=0, 1, 2, \cdots. \tag{3.2}$$

To prove (3.2) we will use Theorem 2.4 of [2], which states that the inequality

$$P_n^\alpha(x)P_{n+1}^\beta(x)-P_{n+1}^\alpha(x)P_n^\beta(x)>0 \tag{3.3}$$

holds for $0<x<1$ if $\frac{1}{2}<\alpha<\beta\leqq\alpha+2$.

It follows from the expansion

$$P_n^\lambda(x)=\sum_{k=0}^{[n/2]}\frac{(-1)^k(\lambda)_{n-k}}{k!(n-2k)!}(2x)^{n-2k}$$

that if n is odd then

$$P_n^1(x)P_{n+1}^\lambda(x)-P_{n+1}^1(x)P_n^\lambda(x)=A_n(\lambda)x^3+\cdots$$

where $A_n(\lambda)>0$ for $\lambda>1$. Similarly, if n is even then

$$P_n^1(x)P_{n+1}^\lambda(x)-P_{n+1}^1(x)P_n^\lambda(x)=B_n(\lambda)x+\cdots$$

where $B_n(\lambda)>0$ for $\lambda>1$. Consequently, (3.2) holds in a right-hand neighborhood of 0, and if (3.2) is not true in $(0, \sqrt{3}/2]$, then there exist n, x, λ with $0<x\leqq\sqrt{3}/2$, $3<\lambda<4$ such that

$$P_n^1(x)P_{n+1}^\lambda(x)-P_{n+1}^1(x)P_n^\lambda(x)=0. \tag{3.4}$$

By (3.3) we have, for $n\geqq0$ and $0<x<1$:

(i) $P_n^\lambda(x)P_{n+1}^4(x)-P_{n+1}^\lambda(x)P_n^4(x)>0$, if $2\leqq\lambda<4$,

(ii) $P_n^3(x)P_{n+1}^\lambda(x)-P_{n+1}^3(x)P_n^\lambda(x)>0$ for $3<\lambda<5$,

(iii) $P_n^1(x)P_{n+1}^3(x)-P_{n+1}^1(x)P_n^3(x)>0$.

And by Theorem 1,

(iv) $P_n^1(x)P_{n+1}^4(x)-P_n^4(x)P_{n+1}^1(x)>0$ for $n\geqq0$, $0<x\leqq\sqrt{3}/2$.

Now (3.4) implies that (for some $n\geqq0$, x and λ with $0<x\leqq\sqrt{3}/2$ and $3<\lambda<4$),

(v) $P_n^1(x)P_{n+1}^\lambda(x)=P_{n+1}^1(x)P_n^\lambda(x)$.

It follows from (v) that (for the same n, λ, x)

(vi) $$P_{n+1}^1(x)[P_n^\lambda(x)P_{n+1}^N(x)-P_{n+1}^\lambda(x)P_n^N(x)] = P_{n+1}^\lambda(x)[P_n^1(x)P_{n+1}^N(x)-P_{n+1}^1(x)P_n^N(x)]$$

for $N=3$ and for $N=4$. But for $N=3$ the quantities in square brackets on either side of (vi) have opposite signs, by (ii) and (iii); for $N=4$ they have the same sign, by (i) and (iv).

It follows that if (v) holds, then (for the same n, x, λ) we have

$$P_{n+1}^{1}(x) = P_{n+1}^{\lambda}(x) = 0.$$

But then, because of (i)-(iv), we would have simultaneously

$$P_n^{\lambda}(x)P_{n+1}^{4}(x) > 0, \qquad P_n^{1}(x)P_{n+1}^{4}(x) > 0,$$

$$P_n^{\lambda}(x)P_{n+1}^{3}(x) < 0, \qquad P_n^{1}(x)P_{n+1}^{3}(x) > 0,$$

which is impossible. Therefore (v) cannot occur.

Acknowledgment. The authors express their appreciation to the unknown referee who so patiently read, corrected, and recorrected this manuscript.

REFERENCES

[1] R. ASKEY, *Some absolutely monotonic functions*, Studia Sci. Math., 9 (1974), pp. 51-56.

[2] R. ASKEY AND G. GASPER, *Positive Jacobi polynomial sums* II, Amer. J. Math., 98 (1976), pp. 709-738.

[3] J. BUSTOZ AND N. SAVAGE, *Inequalities for ultraspherical and Laguerre polynomials*, SIAM J. Math. Anal., 10 (1979), pp. 902-912.

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

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

015

STURM SEQUENCES FOR NONLINEAR EIGENVALUE PROBLEMS*

LEON GREENBERG†

Abstract. An axiomatic treatment of Sturm sequences is presented, and applied to nonlinear eigenvalue problems of the form $\det[A(\lambda)]=0$. Here, $A(\lambda)$ is a symmetric, tridiagonal matrix, whose coefficients $a_{ij}(\lambda)$ satisfy certain monotonicity conditions. This is applied to a finite difference scheme for a Sturm-Liouville problem of the form

$$\begin{cases}(p(x,\lambda)u')'+q(x,\lambda)u=0 & \text{for } a\leqq x\leqq b,\\ \alpha_0(\lambda)u(a)+\beta_0(\lambda)u'(a)=0,\\ \alpha_1(\lambda)u(b)+\beta_1(\lambda)u'(b)=0.\end{cases}$$

The coefficients of this problem are required to satisfy the same kind of monotonicity conditions as those that occur in the Sturm oscillation theorem and that guarantee the existence of eigenvalues.

Key words. Sturm sequence, eigenvalue, nonlinear, symmetric tridiagonal matrix, Sturm-Liouville problem

AMS(MOS) subject classifications. 65H15, 65L15

1. Introduction.

Sturm sequences provide a standard method for finding eigenvalues of symmetric matrices, especially tridiagonal matrices. The case of tridiagonal matrices is of special interest because they arise in the discretization of Sturm-Liouville equations. Also, the Sturm sequence can be calculated recursively in this case. In this paper, we will show that Sturm sequences work for certain kinds of nonlinear eigenvalue problems. The method can then be applied to certain Sturm-Liouville equations in which the eigenvalue occurs nonlinearly. As we will see, this includes cases where the eigenvalue occurs in the boundary conditions.

We recall the main facts about Sturm sequences (see Weber [15, Chap. 8], Givens [3]). Consider a symmetric, tridiagonal matrix:

$$A=\begin{bmatrix} a_1 & b_1 & & & & & \\ b_1 & a_2 & b_2 & & & & \\ & b_2 & a_3 & b_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & b_{n-2} & a_{n-1} & b_{n-1} & \\ & & & & b_{n-1} & a_n & \end{bmatrix}, \tag{1.1}$$

in which $b_i\neq 0$ $(1\leqq i\leqq n-1)$. For a fixed choice of λ_0, let $S_i(\lambda_0)$ be the ith principal minor of $A-\lambda_0 I$. Thus, for $1\leqq i\leqq n$,

$$S_i(\lambda_0)=\begin{vmatrix} a_1-\lambda_0 & b_1 & & & & \\ b_1 & a_2-\lambda_0 & b_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & b_{i-2} & a_{i-1}-\lambda_0 & b_{i-1} & \\ & & & b_{i-1} & a_i-\lambda_0 & \end{vmatrix}. \tag{1.2}$$

We also define $S_0(\lambda_0)=1$. The (finite) sequence $S_0(\lambda_0), S_1(\lambda_0), \cdots, S_n(\lambda_0)$ is the Sturm sequence for the matrix A at λ_0. Let $c(\lambda_0)$ be the number of sign changes in the Sturm

* Received by the editors March 30, 1987; accepted for publication (in revised form) April 6, 1988.
† Mathematics Department, University of Maryland, College Park, Maryland 20742.

sequence (i.e., the number of indices i, such that $S_{i-1}(\lambda_0)S_i(\lambda_0)<0$). If some of the $S_i(\lambda_0)$ are zero, then all zeros are omitted from the sequence, and $c(\lambda_0)$ is the number of sign changes in the remaining sequence. The following three theorems are the main facts about Sturm sequences.

THEOREM A (Eigenvalue count). *$c(\lambda_0)$ equals the number of eigenvalues of A that are less than λ_0.*

THEOREM B (Interlacing theorem). *For $1 \leqq i \leqq n$, the zeros of $S_i(\lambda)$ and $S_{i-1}(\lambda)$ are interlaced (i.e., each function has a unique zero between two consecutive zeros of the other function).*

THEOREM C (Simple eigenvalues). *$S_n(\lambda)$ has exactly n different zeros.*

We also will need to refer to the following weak form of Theorem A.

THEOREM A (Weak form). *If $\lambda' < \lambda''$, then the number of eigenvalues in the interval $[\lambda', \lambda'')$ is $c(\lambda'') - c(\lambda')$.*

Perhaps the most important of these theorems is Theorem A, because it provides a method for calculating eigenvalues. If the eigenvalues of A are $\lambda_1 < \lambda_2 < \cdots < \lambda_n$, we can approximate the kth eigenvalue λ_k by using the bisection method with Theorem A. By bisection, we find a nested sequence of intervals $[\lambda', \lambda'']$ such that $c(\lambda') = k-1$ and $c(\lambda'') = k$. Theorem A guarantees that $\lambda' \leqq \lambda_k < \lambda''$.

We will show that, under certain conditions, Theorems A, B, and C remain true when $A - \lambda I$ is replaced by a more general (symmetric, tridiagonal) matrix $A(\lambda)$, whose coefficients $a_{ij}(\lambda)$ may be nonlinear functions of λ. (Here, an eigenvalue is a zero of $\det[A(\lambda)]$.) In § 2, Sturm sequences are considered from a general, axiomatic point of view. Versions of Theorems A, B, and C are proved in this context. In § 3, the results of § 2 are applied to nonlinear eigenvalue problems for tridiagonal matrices. In § 4, we consider a discretization of a Sturm–Liouville problem. Both the continuous and discrete problems are nonlinear eigenvalue problems. It is shown that Sturm sequences may be applied to the discrete problem, if the coefficients of the continuous problem satisfy the monotonicity conditions in the Sturm oscillation theorem.

Historical note. Sturm sequences first made their appearance in 1829, when Sturm announced his results in [12] and [13]. The announcement [12] was followed by a memoir [14] in 1835, where Sturm proves his well-known theorem relating the number of zeros of a polynomial in an interval to the number of sign changes in a Sturm sequence. The connection of Sturm sequences to eigenvalue problems first appeared in Sturm's announcement [13]. Here, Sturm states versions of Theorems A and B for a matrix of the form $A(\lambda) = K + \lambda G$, where K and G are symmetric matrices (not necessarily tridiagonal), and G is positive definite. Unfortunately, this was not followed by published proofs. The first proof of Theorem A seems to be due to Jacobi [7], in 1857. Here, Jacobi proves the formula

$$\sum_{i,j=1}^{n} a_{ij}x_i x_j = \frac{y_1^2}{S_1} + \sum_{k=2}^{n} \frac{y_k^2}{S_{k-1}S_k},$$

where y_k is a linear, homogeneous function of $x_k, x_{k+1}, \cdots, x_n$, and S_k is the kth principal minor of the symmetric matrix $A = (a_{ij})$. This formula implies that the number of negative eigenvalues of A is equal to the number of sign changes in the sequence $1, S_1, S_2, \cdots, S_n$. Theorem A follows from this fact. The algebra book [10], published by Salmon in 1859, includes a proof of Theorem A, using Sturm sequences. (The proof can be found in § 156 of the first edition. It occurs in § 46 of the fifth edition, which has been published as a Chelsea reprint.) An extensive discussion of Sturm sequences can be found in Chapter 8 of Weber's algebra book (1895) [15], including proofs of

Theorems A and B, and further developments by Hurwitz and Kronecker. Bôcher (1912) [1] has written an excellent review of Sturm's work, in which he discusses Sturm's motivation from problems in mechanics and heat flow.

2. Sturm sequences. In this section, we give an axiomatic development of Sturm sequences. Similar axiomatic treatments can be found in Weber [15], Isaacson and Keller [6], and Stoer and Bulirsch [11], where the weak form of Theorem A (in § 1) is proved within the axiomatic framework. We will develop the properties needed to yield analogues of Theorems A, B, and C. In § 3, this will be applied to nonlinear eigenvalue problems for symmetric, tridiagonal matrices.

DEFINITION 2.1. Let $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ be continuous functions on an interval $\Lambda_1 < \lambda < \Lambda_2$. (The possibilities $\Lambda_1 = -\infty$, $\Lambda_2 = \infty$ are included.) We say that the (finite) sequence $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ has the *Sturm property* if:

(a) $S_0(\lambda)$ has no zeros in (Λ_1, Λ_2).

(b) For $1 \leqq i \leqq n$, the set of zeros of $S_i(\lambda)$ is discrete (i.e., the set of zeros has no limit point in (Λ_1, Λ_2); equivalently, each function $S_i(\lambda)$ has only isolated zeros).

(c) For $1 \leqq i \leqq n-1$, if $S_i(\lambda_0) = 0$, then $S_{i-1}(\lambda_0) S_{i+1}(\lambda_0) < 0$.

(d) If $S_n(\lambda_0) = 0$, then for sufficiently small $\varepsilon_1, \varepsilon_2 > 0$, $S_{n-1}(\lambda_0)[S_n(\lambda_0 + \varepsilon_2) - S_n(\lambda_0 - \varepsilon_1)] < 0$.

Remarks. (1) Conditions (c) and (d) imply that for each λ_0 in (Λ_1, Λ_2), the sequence of numbers $S_0(\lambda_0), S_1(\lambda_0), \cdots, S_n(\lambda_0)$ does not contain two consecutive zeros.

(2) Condition (d) implies that if $S_n(\lambda_0) = 0$, then $S_n(\lambda)$ changes sign as λ passes through λ_0. This need not be true for $S_i(\lambda) (i < n)$. For example, the sequence $1, -\lambda^2, \lambda - 1$ has the Sturm property, but $S_1(\lambda) = -\lambda^2$ does not change sign as λ passes through 0.

(3) Condition (d) is equivalent to two other conditions: If $S_n(\lambda_0) = 0$, then for sufficiently small $\varepsilon > 0$,

(d1) $S_n(\lambda_0 - \varepsilon) S_n(\lambda_0 + \varepsilon) < 0$, and

(d2) $S_{n-1}(\lambda_0)[S_n(\lambda_0 + \varepsilon) - S_n(\lambda_0 - \varepsilon)] < 0$.

(4) In case $S_n(\lambda)$ is differentiable, condition (d) follows from the following condition: If $S_n(\lambda_0) = 0$, then $S_{n-1}(\lambda_0) S_n'(\lambda_0) < 0$.

(5) Condition (d) implies that if $S_n(\lambda_0) = 0$, then for sufficiently small $\varepsilon > 0$, $S_{n-1}(\lambda_0 - \varepsilon) S_n(\lambda_0 - \varepsilon) > 0$ and $S_{n-1}(\lambda_0 + \varepsilon) S_n(\lambda_0 + \varepsilon) < 0$. Thus, a sign change is generated between $S_{n-1}(\lambda)$ and $S_n(\lambda)$, when λ passes through λ_0 from left to right.

DEFINITION 2.2. Let $\alpha_0, \alpha_1, \cdots, \alpha_n$ be a sequence of (real) numbers. $c(\alpha_0, \alpha_1, \cdots, \alpha_n)$ denotes the number of sign changes in $\alpha_0, \alpha_1, \cdots, \alpha_n$ after the zero terms have been omitted. Corresponding to a sequence of functions $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$, the number of sign changes $c(S_0(\lambda_0), S_1(\lambda_0), \cdots, S_n(\lambda_0))$ will usually be denoted simply by $c(\lambda_0)$.

The following theorem is analogous to the weak form of Theorem A in § 1.

THEOREM 2.1. *Suppose that the sequence $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$, has the Sturm property on (Λ_1, Λ_2). Let $\lambda' < \lambda''$ be numbers in (Λ_1, Λ_2). Then $S_n(\lambda)$ has exactly $c(\lambda'') - c(\lambda')$ different zeros in the interval $[\lambda', \lambda'')$.*

Proof. Consider what happens to $c(\lambda)$ as λ increases from λ' to λ''. The interval $[\lambda', \lambda'']$ contains only a finite number of zeros (perhaps none) of any of the functions $S_i(\lambda), 1 \leqq i \leqq n$. (Recall that $S_0(\lambda)$ has no zeros.) $c(\lambda)$ does not change in any subinterval that contains no zeros of $S_i(\lambda), 1 \leqq i \leqq n$.

Suppose that λ_0 is a zero of $S_i(\lambda)$, where $i < n$. By condition (c) in Definition 2.1, $S_{i-1}(\lambda)$ and $S_{i+1}(\lambda)$ have opposite signs in an interval $(\lambda_0 - \varepsilon, \lambda_0 + \varepsilon)$. We suppose that ε is small enough so that the functions $S_j(\lambda), 1 \leqq j \leqq n$, have no zeros other than λ_0 in

$(\lambda_0-\varepsilon, \lambda_0+\varepsilon)$. The subsequence $S_{i-1}(\lambda)$, $S_i(\lambda)$, $S_{i+1}(\lambda)$ has exactly one sign change, for all λ in this interval. (For $\lambda=\lambda_0$, $S_i(\lambda_0)$ is omitted from the subsequence.) The same is true for any other j, $1\leqq j\leqq n-1$, if $S_j(\lambda_0)=0$. Therefore, if $S_n(\lambda_0)\neq 0$, then $c(\lambda)$ does not change in the interval $(\lambda_0-\varepsilon, \lambda_0+\varepsilon)$.

Now suppose that λ_0 is a zero of $S_n(\lambda)$. As we mentioned before (Remark (5)), condition (d) in Definition 2.1 implies that, for small $\varepsilon>0$, $S_{n-1}(\lambda_0-\varepsilon)S_n(\lambda_0-\varepsilon)>0$ and $S_{n-1}(\lambda_0+\varepsilon)S_n(\lambda_0+\varepsilon)<0$. We suppose that λ_0 is the only zero of $S_n(\lambda)$ in $(\lambda_0-\varepsilon, \lambda_0+\varepsilon)$, and $S_{n-1}(\lambda)$ has no zeros in this interval. Then $S_{n-1}(\lambda)S_n(\lambda)>0$ for $\lambda_0-\varepsilon\leqq\lambda<\lambda_0$, and $S_{n-1}(\lambda)S_n(\lambda)<0$ for $\lambda_0<\lambda\leqq\lambda_0+\varepsilon$. Therefore $c(\lambda)=c(\lambda_0)$ for $\lambda_0-\varepsilon\leqq\lambda<\lambda_0$, and $c(\lambda)=c(\lambda_0)+1$ for $\lambda_0<\lambda\leqq\lambda_0+\varepsilon$. In other words, $c(\lambda)$ increases by 1 as λ passes through λ_0 from left to right.

Now consider the interval $[\lambda', \lambda'']$ again. If λ' is a zero of $S_n(\lambda)$, then for small ε, $c(\lambda'+\varepsilon)=c(\lambda')+1$. Thus $c(\lambda)$ "counts" λ'. Similarly, $c(\lambda)$ counts the zeros of S_n in the interior of $[\lambda', \lambda'']$. If λ'' is a zero of $S_n(\lambda)$, then for small ε, $c(\lambda''-\varepsilon)=c(\lambda'')$, so $c(\lambda)$ does not count λ''. □

COROLLARY 2.1. *If the sequence $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ has the Sturm property on (Λ_1, Λ_2), then $S_n(\lambda)$ has at most n different zeros in (Λ_1, Λ_2).*

Proof. For any λ in (Λ_1, Λ_2), $c(\lambda)\leqq n$. □

We now mention a theorem analogous to the strong form of Theorem A in § 1.

THEOREM 2.2. *Let $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ have the Sturm property on (Λ_1, Λ_2), and suppose that $S_0(\lambda)>0$ on (Λ_1, Λ_2). The following are equivalent:*

(1) *There is a number λ_+ in (Λ_1, Λ_2), such that $S_i(\lambda_+)\geqq 0$, for $1\leqq i\leqq n$.*

(2) *For any λ_0 in (Λ_1, Λ_2), $c(\lambda_0)$ equals the number of zeros of $S_n(\lambda)$ in the interval (Λ_1, λ_0).*

Proof. Suppose (1) is true. Then $c(\lambda_+)=0$. Note that Theorem 2.1 implies that $c(\lambda)$ is a nondecreasing function. Therefore $c(\lambda)=0$ for $\Lambda_1<\lambda\leqq\lambda_+$. The theorem also implies that $S_n(\lambda)$ has no zeros in (Λ_1, λ_+).

If $\Lambda_1<\lambda_0\leqq\lambda_+$, then the number of zeros of $S_n(\lambda)$ in (Λ_1, λ_0) is $0=c(\lambda_0)$. If $\lambda_+<\lambda_0<\Lambda_2$, then the zeros of $S_n(\lambda)$ in (Λ_1, λ_0) lie in the subinterval $[\lambda_+, \lambda_0)$. Thus the number of zeros in (Λ_1, λ_0) equals the number of zeros in $[\lambda_+, \lambda_0)$, which equals $c(\lambda_0)-c(\lambda_+)=c(\lambda_0)$. This proves (2).

Now suppose that (2) is true. By Corollary 2.1, $S_n(\lambda)$ has only a finite number of zeros in (Λ_1, Λ_2). Let λ_+ be a number in (Λ_1, Λ_2), which lies to the left of all zeros of $S_n(\lambda)$. By (2), $c(\lambda_+)=0$. Therefore the sequence $S_0(\lambda_+), S_1(\lambda_+), \cdots, S_n(\lambda_+)$ has no sign changes. Therefore $S_i(\lambda_+)\geqq 0$, for $1\leqq i\leqq n$. (Necessarily, $S_n(\lambda_+)>0$.) This proves (1). □

Remark. We have seen that the Sturm property implies the weak form of Theorem A. However, it does not imply Theorem B. For example, the sequence $S_0(\lambda)=1$, $S_1(\lambda)=1-\lambda^2$, $S_2(\lambda)=\lambda-2$ has the Sturm property, but the zeros of S_1 and S_2 are not interlaced. In the following definition, we consider an additional property, which will imply Theorem B.

DEFINITION 2.3. A *Sturm sequence* is a sequence of functions $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$, such that for $1\leqq i\leqq n$, the initial subsequence $S_0(\lambda), S_1(\lambda), \cdots, S_i(\lambda)$ has the Sturm property. In other words, a Sturm sequence satisfies all conditions in Definition 2.1, except that condition (d) is replaced by the stronger condition:

(d′) For $1\leqq i\leqq n$, if $S_i(\lambda_0)=0$, then for sufficiently small $\varepsilon_1, \varepsilon_2>0$, $S_{i-1}(\lambda_0)[S_i(\lambda_0+\varepsilon_2)-S_i(\lambda_0-\varepsilon_1)]<0$.

Remark. Condition (d′) implies that for $1\leqq i\leqq n$, if $S_i(\lambda_0)=0$, then $S_i(\lambda)$ changes sign as λ passes through λ_0.

THEOREM 2.3. *Suppose that* $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ *is a Sturm sequence. Then, for* $1 \leqq i \leqq n$, *the zeros of* $S_i(\lambda)$ *and* $S_{i-1}(\lambda)$ *are interlaced.*

Proof. We use induction on i. The statement is true for $i=1$, since $S_0(\lambda)$ has no zeros, and by Corollary 2.1, $S_1(\lambda)$ has at most one zero. Suppose the statement is true for $i \leqq k$. We will prove it for $i=k+1$.

Let $\lambda_1 < \lambda_2$ be consecutive zeros of $S_{k+1}(\lambda)$. We will show that $S_k(\lambda)$ has a zero between λ_1 and λ_2. ($S_k(\lambda)$ cannot have more than one zero in (λ_1, λ_2), because we will later show that $S_{k+1}(\lambda)$ has a zero between two zeros of $S_k(\lambda)$.) Condition (d') in Definition 2.3 implies that for sufficiently small $\varepsilon > 0$, and for $j=1, 2$, $S_k(\lambda_j - \varepsilon) S_{k+1}(\lambda_j - \varepsilon) > 0$ and $S_k(\lambda_j + \varepsilon) S_{k+1}(\lambda_j + \varepsilon) < 0$. Thus, $S_k(\lambda_1 + \varepsilon)$ and $S_{k+1}(\lambda_1 + \varepsilon)$ have opposite signs, while $S_k(\lambda_2 - \varepsilon)$ and $S_{k+1}(\lambda_2 - \varepsilon)$ have the same sign. Since λ_1 and λ_2 are consecutive zeros, $S_{k+1}(\lambda)$ does not change sign in the interval (λ_1, λ_2). Therefore, for small $\varepsilon > 0$, $S_k(\lambda_1 + \varepsilon)$ and $S_k(\lambda_2 - \varepsilon)$ have opposite signs, so $S_k(\lambda)$ has a zero in (λ_1, λ_2).

Now let $\lambda_1 < \lambda_2$ be consecutive zeros of $S_k(\lambda)$. We will show that $S_{k+1}(\lambda)$ has a zero between λ_1 and λ_2. (As before, this zero must be unique.) By condition (c) in Definition 2.1, $S_{k-1}(\lambda_j) S_{k+1}(\lambda_j) < 0$, for $j=1, 2$. By inductive hypothesis, $S_{k-1}(\lambda)$ has a unique zero λ_0 between λ_1 and λ_2. $S_{k-1}(\lambda)$ changes sign at any zero, so $S_{k-1}(\lambda_1)$ and $S_{k-1}(\lambda_2)$ have opposite signs. Since $S_{k-1}(\lambda_j) S_{k+1}(\lambda_j) < 0$, for $j=1, 2$, it follows that $S_{k+1}(\lambda_1)$ and $S_{k+1}(\lambda_2)$ have opposite signs. Therefore $S_{k+1}(\lambda)$ has a zero between λ_1 and λ_2. □

Remark. If $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ is a Sturm sequence on (Λ_1, Λ_2), $S_n(\lambda)$ need not have n zeros in (Λ_1, Λ_2), since we may have omitted some of the zeros by taking the interval too small. Even if the interval is $(-\infty, \infty)$, $S_n(\lambda)$ may not have n zeros. An example of this is $S_0(\lambda)=1$, $S_1(\lambda)=-\lambda-1$, $S_2(\lambda)=\lambda-1$. The following theorem (analogous to Theorem C in § 1) gives necessary and sufficient conditions for $S_n(\lambda)$ to have n zeros.

THEOREM 2.4. *Suppose that the sequence* $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ *is a Sturm sequence on* (Λ_1, Λ_2), *and that* $S_0(\lambda) > 0$, *for* $\Lambda_1 < \lambda < \Lambda_2$. *Then* $S_n(\lambda)$ *has exactly n different zeros in* (Λ_1, Λ_2) *if and only if there are numbers* $\lambda_+ < \lambda_-$ *in* (Λ_1, Λ_2), *such that, for* $1 \leqq i \leqq n$,

(1) $S_i(\lambda_+) \geqq 0$, *and*

(2) *the sign of* $S_i(\lambda_-)$ *is* $(-1)^i$.

If these conditions are satisfied, then $S_i(\lambda)$ *has exactly i different zeros in* (Λ_1, Λ_2), *for* $1 \leqq i \leqq n$.

Proof. First suppose that $S_n(\lambda)$ has n different zeros in (Λ_1, Λ_2). Since the zeros of $S_{n-1}(\lambda)$ and $S_n(\lambda)$ are interlaced, $S_{n-1}(\lambda)$ has $n-1$ zeros in (Λ_1, Λ_2). Similarly, by induction, we can show that $S_i(\lambda)$ has i zeros in (Λ_1, Λ_2). Let $c_i(\lambda)$ denote the number of sign changes in the sequence $S_0(\lambda), S_1(\lambda), \cdots, S_i(\lambda)$, after the zero terms have been omitted. Let $\lambda_+ < \lambda_-$ be numbers in (Λ_1, Λ_2), such that λ_+ lies to the left of all zeros of all $S_j(\lambda)$, and λ_- lies to the right of these zeros. Then the number of zeros of $S_i(\lambda)$ in (Λ_1, Λ_2) is $i = c_i(\lambda_-) - c_i(\lambda_+)$. Since $0 \leqq c_i(\lambda) \leqq i$, this implies that $c_i(\lambda_+) = 0$ and $c_i(\lambda_-) = i$. Therefore, $S_i(\lambda_+) \geqq 0$ and $S_i(\lambda_-)$ has sign $(-1)^i$. This proves (1) and (2).

Now suppose conditions (1) and (2) are satisfied. Then $c_i(\lambda_+) = 0$ and $c_i(\lambda_-) = i$. By Theorem 2.1, $S_i(\lambda)$ has $c_i(\lambda_-) - c_i(\lambda_+) = i$ zeros in the interval $[\lambda_+, \lambda_-)$. By Corollary 2.1, $S_i(\lambda)$ has at most i zeros. Therefore it has exactly i zeros in (Λ_1, Λ_2). □

3. Tridiagonal matrices. We will apply the results of § 2 to symmetric, tridiagonal matrices $A(\lambda)$. In each case, the eigenvalues are the zeros of $\det[A(\lambda)]$, $S_i(\lambda)$ (for $1 \leqq i \leqq n$) is the ith principal minor of $A(\lambda)$, and $S_0(\lambda) = 1$. We will begin with matrices

of the form

$$(3.1)\quad A(\lambda)=\begin{bmatrix} (b_0+b_1-a_1) & -b_1 & & & & & \\ -b_1 & (b_1+b_2-a_2) & -b_2 & & & & \\ & -b_2 & (b_2+b_3-a_3) & -b_3 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -b_{n-2} & (b_{n-2}+b_{n-1}-a_{n-1}) & -b_{n-1} & \\ & & & & -b_{n-1} & (b_{n-1}+b_n-a_n) \end{bmatrix},$$

where the $a_i=a_i(\lambda)$ and $b_j=b_j(\lambda)$ are continuous functions on an interval (Λ_1,Λ_2). (The possibilities $\Lambda_1=-\infty$, $\Lambda_2=\infty$ are included.)

The following is a well-known recursion relation for the $S_i(\lambda)$.

LEMMA 3.1. *$S_{i+1}=(b_i+b_{i+1}-a_{i+1})S_i-b_i^2S_{i-1}$, for $1\leqq i\leqq n-1$.*

Proof. This is the cofactor expansion of the determinant

$$S_{i+1}=\begin{vmatrix} (b_0+b_1-a_1) & -b_1 & & & & & \\ -b_1 & (b_1+b_2-a_2) & -b_2 & & & & \\ & -b_2 & (b_2+b_3-a_3) & -b_3 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -b_{i-1} & (b_{i-1}+b_i-a_i) & -b_i & \\ & & & & -b_i & (b_i+b_{i+1}-a_{i+1}) \end{vmatrix}$$

by the last row. □

Remark. If $b_i(\lambda)$ has no zeros (for $1\leqq i\leqq n-1$), then Lemma 3.1 implies that two consecutive functions S_i, S_{i+1} cannot have a common zero. If $S_i(\lambda_0)=0=S_{i+1}(\lambda_0)$, then the recursion relation implies that $S_{i-1}(\lambda_0)=0$. This in turn implies $0=S_{i-1}(\lambda_0)=S_{i-2}(\lambda_0)=S_{i-3}(\lambda_0)=\cdots=S_0(\lambda_0)$, which contradicts $S_0(\lambda)=1$. The recursion relation also implies that if $S_i(\lambda_0)=0$, then $S_{i+1}(\lambda_0)S_{i-1}(\lambda_0)\leqq 0$. Since two consecutive functions have no common zero, $S_{i+1}(\lambda_0)S_{i-1}(\lambda_0)<0$. This is condition (c) in Definition 2.1.

LEMMA 3.2. *Suppose that*

(1) *$a_i(\lambda)$ is strictly increasing, for $1\leqq i\leqq n$;*

(2) *$b_j(\lambda)$ is nonincreasing, for $0\leqq j\leqq n$;*

(3) *$b_i(\lambda)$ has no zeros, for $1\leqq i\leqq n-1$.*

Then,

(a) *$S_i(\lambda)$ has only isolated zeros, for $1\leqq i\leqq n$;*

(b) *$b_i-b_i^2(S_{i-1}/S_i)$ is a strictly decreasing function on any interval containing no zeros of $S_i(\lambda)$, for $1\leqq i\leqq n-1$.*

Proof. We will prove (a) and (b) simultaneously by induction on i. Note that $i=n$ is permitted in (a), while $i\leqq n-1$ in (b). The reason for the restriction in (b) is that $b_n(\lambda)$ is allowed to have zeros or even to be identically zero.

For $i=1$, $S_1=b_0+b_1-a_1$ is strictly decreasing, so it can have at most one zero. We must show that $b_1-b_1^2(S_0/S_1)$ is strictly decreasing on any interval that contains no zeros of S_1. We first consider an interval that has no zeros of S_1 or b_0-a_1:

$$b_1-b_1^2\frac{S_0}{S_1}=b_1-\frac{b_1^2}{b_0+b_1-a_1}=\frac{b_1(b_0-a_1)}{b_1+(b_0-a_1)}.$$

This shows that $b_1-b_1^2(S_0/S_1)$ has no zeros in the given interval, and

$$\left[b_1-b_1^2\frac{S_0}{S_1}\right]^{-1}=\frac{1}{b_0-a_1}+\frac{1}{b_1}.$$

Since b_0-a_1 is strictly decreasing and b_1 is nonincreasing, this shows that $[b_1-b_1^2(S_0/S_1)]^{-1}$ is strictly increasing, so $b_1-b_1^2(S_0/S_1)$ is strictly decreasing in the given interval.

Since $b_0 - a_1$ is strictly decreasing, it has at most one zero λ_0. Suppose that (α, β) is an interval that contains λ_0, but no zeros of S_1. By the previous paragraph, $b_1 - b_1^2(S_0/S_1)$ is strictly decreasing in (α, λ_0) and (λ_0, β). By continuity, it is strictly decreasing in (α, β). This concludes the inductive step $i = 1$.

Now suppose that statements (a) and (b) are true for $i \leqq k$; we will prove them for $i = k+1$. First we consider (a). By Lemma 3.1,

$$S_{k+1} = (b_k + b_{k+1} - a_{k+1})S_k - b_k^2 S_{k-1}, \quad \text{or}$$

$$S_{k+1} = S_k\left[(b_{k+1} - a_{k+1}) + \left(b_k - b_k^2 \frac{S_{k-1}}{S_k}\right)\right]. \tag{3.2}$$

As we mentioned in the remark following Lemma 3.1, S_{k+1} and S_k cannot have a common zero. By induction, the set of zeros of S_k is discrete, and $b_k - b_k^2(S_{k-1}/S_k)$ is strictly decreasing in any interval that contains no zeros of S_k. Therefore the function $(b_{k+1} - a_{k+1}) + (b_k - b_k^2(S_{k-1}/S_k))$ is strictly decreasing between any two consecutive zeros of S_k. Equation (3.2) shows that S_{k+1} can have at most one zero between any two consecutive zeros of S_k. This proves statement (a) for $i = k+1$.

Now we consider statement (b). Here we assume that $k+1 \leqq n-1$. Consequently, $b_{k+1}(\lambda)$ has no zeros. From the recursion relation

$$S_{k+1} = (b_k + b_{k+1} - a_{k+1})S_k - b_k^2 S_{k-1},$$

we obtain

$$\frac{S_{k+1}}{S_k} = (b_k + b_{k+1} - a_{k+1}) - b_k^2 \frac{S_{k-1}}{S_k},$$

$$\frac{S_k}{S_{k+1}} = \frac{1}{(b_k + b_{k+1} - a_{k+1}) - b_k^2(S_{k-1}/S_k)},$$

$$b_{k+1} - b_{k+1}^2 \frac{S_k}{S_{k+1}} = b_{k+1} - \frac{b_{k+1}^2}{(b_k + b_{k+1} - a_{k+1}) - b_k^2(S_{k-1}/S_k)},$$

and finally

$$b_{k+1} - b_{k+1}^2 \frac{S_k}{S_{k+1}} = \frac{b_{k+1}\left[\left(b_k - b_k^2 \dfrac{S_{k-1}}{S_k}\right) - a_{k+1}\right]}{b_{k+1} + \left[\left(b_k - b_k^2 \dfrac{S_{k-1}}{S_k}\right) - a_{k+1}\right]}. \tag{3.3}$$

Taking reciprocals in (3.3), we obtain

$$\frac{1}{b_{k+1} - b_{k+1}^2(S_k/S_{k+1})} = \frac{1}{(b_k - b_k^2(S_{k-1}/S_k)) - a_{k+1}} + \frac{1}{b_{k+1}}. \tag{3.4}$$

Let I be an interval that contains no zeros of S_{k+1}, S_k or $(b_k - b_k^2(S_{k-1}/S_k)) - a_{k+1}$. Equation (3.3) shows that I has no zeros of $b_{k+1} - b_{k+1}^2(S_k/S_{k+1})$. Since $(b_k - b_k^2(S_{k-1}/S_k)) - a_{k+1}$ is strictly decreasing in I, and b_{k+1} is nonincreasing, (3.4) shows that $b_{k+1} - b_{k+1}^2(S_k/S_{k+1})$ is strictly decreasing in I. Now let J be an interval that has no zeros of S_{k+1}. The zeros of S_k and $(b_k - b_k^2(S_{k-1}/S_k)) - a_{k+1}$ form a discrete set Z in J, and $b_{k+1} - b_{k+1}^2(S_k/S_{k+1})$ is strictly decreasing in any interval contained in $J - Z$. By continuity, $b_{k+1} - b_{k+1}^2(S_k/S_{k+1})$ is strictly decreasing in J. □

Remark. Conclusion (b) in Lemma 3.2 will be vital to the proof of Theorem 3.1 and all the theorems in this section that follow from it. This crucial fact was found by studying the finite difference matrices arising from Sturm-Liouville problems of type (4.1), in § 4. If the Sturm sequence S_i is compared to the finite difference solution u_i of (4.1), then it turns out that $b_i - b_i^2(S_{i-1}/S_i) \approx hp_{i+1/2}(u'_{i+1}/u_{i+1})$. Under the conditions indicated in Theorem D of § 4, one of the Sturm comparison theorems states that $p(x,\lambda)(u'(x,\lambda)/u(x,\lambda))$ is a decreasing function of λ. This fact suggested conclusion (b) in Lemma 3.2. The relationship between the Sturm sequence and the finite difference solution will be treated by Greenberg and Babuška in [4].

THEOREM 3.1. *Suppose that*

(1) $a_i(\lambda)$ *is strictly increasing, for* $1 \leqq i \leqq n$;

(2) $b_j(\lambda)$ *is nonincreasing, for* $0 \leqq j \leqq n$;

(3) $b_j(\lambda)$ *has no zeros, for* $1 \leqq j \leqq n-1$.

Then the sequence $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ *is a Sturm sequence (in the sense of* § 2).

Proof. By Definition 2.3, we must verify (a)-(c) in Definition 2.1 and (d') in Definition 2.3. Condition (a) is obvious, since $S_0(\lambda) = 1$. Condition (b) was proved in Lemma 3.2(a). Condition (c) follows from the remark after Lemma 3.1. It remains to verify the condition:

(d') For $1 \leqq i \leqq n$, if $S_i(\lambda_0) = 0$, then for sufficiently small $\varepsilon_1, \varepsilon_2 > 0$,

$$S_{i-1}(\lambda_0)[S_i(\lambda_0+\varepsilon_2) - S_i(\lambda_0-\varepsilon_1)] < 0.$$

To verify the ith statement (d'), we will need the $(i-1)$st statement in Lemma 3.2(b), which is valid for $1 \leqq i-1 \leqq n-1$. Because of this, we need a separate calculation for $i=1$ in (d'):

$$S_0(\lambda_0)[S_1(\lambda_0+\varepsilon_2) - S_1(\lambda_0-\varepsilon_1)] < 0.$$

This inequality follows from the fact that $S_0(\lambda) = 1$, and $S_1(\lambda) = b_0(\lambda) + b_1(\lambda) - a_1(\lambda)$ is a strictly decreasing function.

Now consider (d') for $2 \leqq i \leqq n$.

By Lemma 3.1,

$$S_i = (b_{i-1} + b_i - a_i)S_{i-1} - b_{i-1}^2 S_{i-2}, \quad \text{or}$$

$$S_i = S_{i-1}\left[\left(b_{i-1} - b_{i-1}^2 \frac{S_{i-2}}{S_{i-1}}\right) + (b_i - a_i)\right]. \tag{3.5}$$

Choose $\varepsilon_1, \varepsilon_2 > 0$ small enough so that the interval $[\lambda_0-\varepsilon_1, \lambda_0+\varepsilon_2]$ contains no zeros of $S_{i-1}(\lambda)$. Thus S_{i-1} does not change sign in this interval. $S_i/S_{i-1} = (b_{i-1} - b_{i-1}^2(S_{i-2}/S_{i-1})) + (b_i - a_i)$ is a strictly decreasing function in $[\lambda_0-\varepsilon_1, \lambda_0+\varepsilon_2]$ (by Lemma 3.2(b) and hypotheses (1) and (2) of this theorem). Therefore S_i/S_{i-1} changes sign from (+) to (−) as λ passes through λ_0 from left to right. The same must be true for $S_{i-1}S_i = S_{i-1}^2(S_i/S_{i-1})$. Thus

$$S_{i-1}(\lambda_0-\varepsilon_1)S_i(\lambda_0-\varepsilon_1) > 0 > S_{i-1}(\lambda_0+\varepsilon_2)S_i(\lambda_0+\varepsilon_2).$$

Since $S_{i-1}(\lambda)$ does not change sign in $[\lambda_0-\varepsilon_1, \lambda_0+\varepsilon_2]$, it follows that

$$S_{i-1}(\lambda_0)S_i(\lambda_0-\varepsilon_1) > 0 > S_{i-1}(\lambda_0)S_i(\lambda_0+\varepsilon_2), \quad \text{or}$$

$$S_{i-1}(\lambda_0)[S_i(\lambda_0+\varepsilon_2) - S_i(\lambda_0-\varepsilon_1)] < 0.$$

This proves condition (d'). □

The following theorem is a weak version of Theorem A (in § 1) for the matrix $A(\lambda)$.

THEOREM 3.2. *Let $A(\lambda)$ be a matrix of type* (3.1), *and suppose that*

(1) *$a_i(\lambda)$ is strictly increasing, for $1 \leqq i \leqq n$;*

(2) *$b_j(\lambda)$ is nonincreasing, for $0 \leqq j \leqq n$;*

(3) *$b_j(\lambda)$ has no zeros, for $1 \leqq j \leqq n-1$.*

Let $\lambda' < \lambda''$ be numbers in (Λ_1, Λ_2). Then $\det[A(\lambda)]$ has exactly $c(\lambda'') - c(\lambda')$ different zeros in the interval $[\lambda', \lambda'')$.

Proof. This follows from Theorems 2.1 and 3.1. □

LEMMA 3.3. *Let A be a matrix of the form* (3.1), *where the $a_i (1 \leqq i \leqq n)$ and $b_j (0 \leqq j \leqq n)$ are constants. Suppose that*

(1) *$b_j \geqq 0$, for $0 \leqq j \leqq n$,*

(2) *$a_i \leqq 0$, for $1 \leqq i \leqq n$.*

Then $\det[A] \geqq 0$.

Proof. A is the matrix of the quadratic form $Q(x) = xAx^T$, where $x = (x_1, x_2, \cdots, x_n)$:

$$Q(x) = \sum_{i=1}^{n} (b_{i-1} + b_i - a_i)x_i^2 - 2\sum_{j=1}^{n-1} b_j x_j x_{j+1}$$

$$= b_0 x_1^2 + b_n x_n^2 - \sum_{i=1}^{n} a_i x_i^2 + \sum_{j=1}^{n-1} b_j (x_j^2 - 2x_j x_{j+1} + x_{j+1}^2)$$

$$= b_0 x_1^2 + b_n x_n^2 - \sum_{i=1}^{n} a_i x_i^2 + \sum_{j=1}^{n-1} b_j (x_j - x_{j+1})^2.$$

Thus $Q(x)$ is positive semidefinite, and therefore $\det[A] \geqq 0$. □

The following theorem is a strong version of Theorem A for the matrix $A(\lambda)$.

THEOREM 3.3. *Let $A(\lambda)$ be a matrix of type* (3.1), *and suppose that*

(1) *$a_i(\lambda)$ is strictly increasing, for $1 \leqq i \leqq n$;*

(2) *$b_j(\lambda)$ is nonincreasing, for $0 \leqq j \leqq n$;*

(3) *$b_j(\lambda) > 0$, for $1 \leqq j \leqq n-1$;*

(4) *There is a number λ_+ in (Λ_1, Λ_2), such that $b_0(\lambda_+) \geqq 0$, $b_n(\lambda_+) \geqq 0$ and $a_i(\lambda_+) \leqq 0$ for $1 \leqq i \leqq n$.*

Then for any λ_0 in (Λ_1, Λ_2), $\det[A(\lambda)]$ has exactly $c(\lambda_0)$ different zeros in the interval (Λ_1, λ_0).

Proof. Lemma 3.3 implies that $S_i(\lambda_+) \geqq 0$, for $0 \leqq i \leqq n$. The theorem now follows from Theorems 2.2 and 3.1. □

The following theorem is analogous to Theorem B (in § 1).

THEOREM 3.4. *Let $A(\lambda)$ be a matrix of type* (3.1), *and suppose that*

(1) *$a_i(\lambda)$ is strictly increasing, for $1 \leqq i \leqq n$;*

(2) *$b_j(\lambda)$ is nonincreasing, for $0 \leqq j \leqq n$;*

(3) *$b_j(\lambda)$ has no zeros, for $1 \leqq j \leqq n-1$.*

Then for $1 \leqq i \leqq n$ the zeros of $S_i(\lambda)$ and $S_{i-1}(\lambda)$ are interlaced.

Proof. This follows from Theorems 2.3 and 3.1. □

LEMMA 3.4. *Suppose that*

(1) *$b_j(\lambda)$ is nonincreasing, for $0 \leqq j \leqq n$;*

(2) *There is a number K, so that $b_j(\lambda) \geqq K$, for $0 \leqq j \leqq n$;*

(3) *$\lim_{\lambda \to \Lambda_2} a_i(\lambda) = \infty$, for $1 \leqq i \leqq n$.*

Then $\lim_{\lambda \to \Lambda_2} S_i(\lambda) = (-1)^i \infty$, for $1 \leqq i \leqq n$.

Proof. Let λ_0 be in (Λ_1, Λ_2). Conditions (1) and (2) imply that $K \leqq b_j(\lambda) \leqq b_j(\lambda_0)$, for $\lambda_0 < \lambda < \Lambda_2$, $0 \leqq j \leqq n$. Thus the $b_j(\lambda)$ are bounded (above and below) in the interval

(λ_0, Λ_2). The determinant $S_i(\lambda)$ can be expanded in the form:

$$S_i(\lambda) = (-1)^i a_1 a_2 \cdots a_i + A_{i-1} + A_{i-2} + \cdots + A_1 + A_0,$$

where

$$A_k = \sum_{1 \leqq j_1 < j_2 < \cdots < j_k \leqq i} B_{j_1 j_2 \cdots j_k} a_{j_1} a_{j_2} \cdots a_{j_k}$$

and $B_{j_1 j_2 \cdots j_k}$ is a sum of products of the $b_j(\lambda)$. The last term A_0 is also a sum of products of the $b_j(\lambda)$. Since the $b_j(\lambda)$ are bounded in the interval (λ_0, Λ_2), so are the coefficients $B_{j_1 j_2 \cdots j_k}$ and A_0. Therefore,

$$\frac{A_k}{a_1 a_2 \cdots a_i} = \sum_{1 \leqq j_1 \leqq j_2 < \cdots < j_k \leqq i} B_{j_1 j_2 \cdots j_k} \frac{a_{j_1} a_{j_2} \cdots a_{j_k}}{a_1 a_2 \cdots a_i} \to 0 \quad \text{as } \lambda \to \Lambda_2.$$

Consequently,

$$\lim_{\lambda \to \Lambda_2} S_i(\lambda) = \lim_{\lambda \to \Lambda_2} a_1 a_2 \cdots a_i \left[(-1)^i + \frac{A_{i-1}}{a_1 a_2 \cdots a_i} + \cdots + \frac{A_0}{a_1 a_2 \cdots a_i} \right] = (-1)^i \infty. \qquad \square$$

The following theorem is analogous to Theorem C (in § 1).

THEOREM 3.5. *Let $A(\lambda)$ be a matrix of type* (3.1), *and suppose that*

(1) *$a_i(\lambda)$ is strictly increasing, for $1 \leqq i \leqq n$;*

(2) *$b_j(\lambda)$ is nonincreasing, for $0 \leqq j \leqq n$;*

(3) *$b_j(\lambda) > 0$, for $1 \leqq j \leqq n-1$, $b_0(\lambda) \geqq 0$, $b_n(\lambda) \geqq 0$;*

(4) *There is a number λ_+ in (Λ_1, Λ_2), such that $a_i(\lambda_+) \leqq 0$, for $1 \leqq i \leqq n$;*

(5) *$\lim_{\lambda \to \Lambda_2} a_i(\lambda) = \infty$, for $1 \leqq i \leqq n$.*

Then $\det[A(\lambda)]$ *has exactly n different zeros in the interval (Λ_1, Λ_2).*

Proof. Lemma 3.4 implies that $\lim_{\lambda \to \Lambda_2} S_i(\lambda) = (-1)^i \infty$, for $1 \leqq i \leqq n$. Lemma 3.3 implies that $S_i(\lambda_+) \geqq 0$, for $0 \leqq i \leqq n$. The theorem now follows from Theorems 2.4 and 3.1. $\square$

Remark. Theorems 3.1–3.5 remain valid if the matrix $A(\lambda)$ in (3.1) is replaced by

$$A(\lambda) = \begin{bmatrix} (b_0 + b_1 - a_1) & b_1 & & & & & \\ b_1 & (b_1 + b_2 - a_2) & b_2 & & & & \\ & b_2 & (b_2 + b_3 - a_3) & b_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & b_{n-2} & (b_{n-2} + b_{n-1} - a_{n-1}) & b_{n-1} & \\ & & & & b_{n-1} & (b_{n-1} + b_n - a_n) \end{bmatrix} \tag{3.6}$$

or

$$A(\lambda) = \begin{bmatrix} (b_0 + b_1 - a_1) & \varepsilon_1 b_1 & & & & & \\ \varepsilon_1 b_1 & (b_1 + b_2 - a_2) & \varepsilon_2 b_2 & & & & \\ & \varepsilon_2 b_2 & (b_2 + b_3 - a_3) & \varepsilon_3 b_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \varepsilon_{n-2} b_{n-2} & (b_{n-2} + b_{n-1} - a_{n-1}) & \varepsilon_{n-1} b_{n-1} & \\ & & & & \varepsilon_{n-1} b_{n-1} & (b_{n-1} + b_n - a_n) \end{bmatrix}, \tag{3.7}$$

where $\varepsilon_i = \pm 1$, for $1 \leqq i \leqq n-1$. This is true because all of these matrices have the same principal minors $S_1(\lambda), S_2(\lambda), \cdots, S_n(\lambda)$.

We will now consider a symmetric tridiagonal matrix of general form:

$$A(\lambda)=\begin{bmatrix} a_1(\lambda) & b_1(\lambda) & & & & & \\ b_1(\lambda) & a_2(\lambda) & b_2(\lambda) & & & & \\ & b_2(\lambda) & a_3(\lambda) & b_3(\lambda) & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & b_{n-2}(\lambda) & a_{n-1}(\lambda) & b_{n-1}(\lambda) & \\ & & & & b_{n-1}(\lambda) & a_n(\lambda) & \end{bmatrix}, \tag{3.8}$$

where the functions $a_i(\lambda)$, $1\leqq i\leqq n$, and $b_j(\lambda)$, $1\leqq j\leqq n-1$, are continuous on an interval (Λ_1, Λ_2). The above matrix can be transformed to a matrix of the form (3.1), which has the same principal minors $S_1(\lambda), S_2(\lambda), \cdots, S_n(\lambda)$. To see this, first multiply the nondiagonal terms $b_j(\lambda)$ by (-1) to obtain

$$\begin{bmatrix} a_1 & -b_1 & & & & & \\ -b_1 & a_2 & -b_2 & & & & \\ & -b_2 & a_3 & -b_3 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -b_{n-2} & a_{n-1} & -b_{n-1} & \\ & & & & -b_{n-1} & a_n & \end{bmatrix}. \tag{3.9}$$

Next, for $1\leqq i\leqq n$, define $\bar{a}_i=b_{i-1}+b_i-a_i$, where $b_0=0=b_n$. Then $a_i=b_{i-1}+b_i-\bar{a}_i$, so the matrix (3.9) has the form

$$\bar{A}(\lambda)=\begin{bmatrix} (b_1-\bar{a}_1) & -b_1 & & & & & \\ -b_1 & (b_1+b_2-\bar{a}_2) & -b_2 & & & & \\ & -b_2 & (b_2+b_3-\bar{a}_3) & -b_3 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -b_{n-2} & (b_{n-2}+b_{n-1}-\bar{a}_{n-1}) & -b_{n-1} & \\ & & & & -b_{n-1} & (b_{n-1}-\bar{a}_n) & \end{bmatrix}. \tag{3.10}$$

$\bar{A}(\lambda)$ has the form (3.1), with $b_0=0=b_n$. Therefore Theorems 3.1-3.5 apply, and we immediately obtain analogous results for the matrix $A(\lambda)$ in (3.8). We will state these results by referring to the following list of conditions. (When any statement involves b_0 or b_n, it will be understood that $b_0=0=b_n$.)

Condition List 3.1.

(1) $b_{i-1}(\lambda)+b_i(\lambda)-a_i(\lambda)$ is strictly increasing, for $1\leqq i\leqq n$.

(2) $b_j(\lambda)$ is nonincreasing, for $1\leqq j\leqq n-1$.

(3) $b_j(\lambda)$ has no zeros, for $1\leqq j\leqq n-1$.

(4) $b_j(\lambda)>0$, for $1\leqq j\leqq n-1$.

(5) There is a number λ_+ in (Λ_1, Λ_2), such that $b_{i-1}(\lambda_+)+b_i(\lambda_+)-a_i(\lambda_+)\leqq 0$, for $1\leqq i\leqq n$.

(6) $\lim_{\lambda\to\Lambda_2}[b_{i-1}(\lambda)+b_i(\lambda)-a_i(\lambda)]=\infty$, for $1\leqq i\leqq n$.

We continue to use our standard notation: $S_i(\lambda)$ is the ith principal minor of the matrix $A(\lambda)$ in (3.8); $S_0(\lambda)=1$; $c(\lambda)$ is the number of sign changes in the sequence $S_0(\lambda), S_1(\lambda), \cdots, S_n(\lambda)$ after the zero terms have been omitted.

THEOREM 3.6. *Let $A(\lambda)$ be a matrix of type* (3.8), *which satisfies conditions* (1), (2), *and* (3) *in Condition List* 3.1. *Let $\lambda'<\lambda''$ be numbers in (Λ_1, Λ_2). Then* $\det[A(\lambda)]$ *has exactly $c(\lambda'')-c(\lambda')$ different zeros in the interval $[\lambda', \lambda'')$.*

THEOREM 3.7. *Let $A(\lambda)$ be a matrix of type* (3.8), *which satisfies conditions* (1), (2), (4), *and* (5) *in Condition List* 3.1. *Then for any λ_0 in (Λ_1, Λ_2), $\det[A(\lambda)]$ has exactly $c(\lambda_0)$ different zeros in the interval (Λ_1, λ_0).*

THEOREM 3.8. *Let $A(\lambda)$ be a matrix of type* (3.8), *which satisfies conditions* (1), (2), *and* (3) *in Condition List* 3.1. *Then for $1 \leqq i \leqq n$, the zeros of $S_i(\lambda)$ and $S_{i-1}(\lambda)$ are interlaced.*

THEOREM 3.9. *Let $A(\lambda)$ be a matrix of type* (3.8), *which satisfies conditions* (1), (2), (4), (5), *and* (6) *in Condition List* 3.1. *Then $\det[A(\lambda)]$ has exactly n different zeros in the interval (Λ_1, Λ_2).*

We mention a special case of (3.8), in which the functions $b_j(\lambda)$ are constant, $1 \leqq j \leqq n-1$.

THEOREM 3.10. *Let $A(\lambda)$ be a matrix of type* (3.8) *in which the functions $b_j(\lambda)$ are nonzero constants for $1 \leqq j \leqq n-1$. Suppose that, for $1 \leqq i \leqq n$,*

(i) *$a_i(\lambda)$ is strictly decreasing*;

(ii) $\lim_{\lambda \to \Lambda_1} a_i(\lambda) = \infty$;

(iii) $\lim_{\lambda \to \Lambda_2} a_i(\lambda) = -\infty$.

Then,

(a) *For any λ_0 in (Λ_1, Λ_2), $\det[A(\lambda)]$ has exactly $c(\lambda_0)$ different zeros in the interval (Λ_1, λ_0)*;

(b) *The zeros of $S_i(\lambda)$ and $S_{i-1}(\lambda)$ are interlaced, for $1 \leqq i \leqq n$*;

(c) *$\det[A(\lambda)]$ has exactly n different zeros in (Λ_1, Λ_2).*

Proof. (a) By Theorem 3.7, it suffices to verify conditions (1), (2), (4), (5) in Condition List 3.1. Condition (1) follows from (i). Condition (2) is true because the $b_j(\lambda)$ are constants. Condition (5) follows from (ii). Condition (4) requires that $b_j > 0$. Although this might not be true, the principal minors of $A(\lambda)$ are not changed if b_j is replaced by $-b_j$. Therefore, we may assume that $b_j > 0$.

(b) By Theorem 3.8, we need to verify conditions (1), (2), and (3) in Condition List 3.1. We have already verified (1) and (2). Since the b_j are assumed to be nonzero, (3) is also true.

(c) By Theorem 3.9, (c) follows from conditions (1), (2), (4)-(6) in Condition List 3.1. The first four of these conditions were verified in the proof of (a). Condition (6) follows from (iii). □

Remarks. We can obtain results similar to Theorems 3.2-3.5 for the matrix (3.1), if the $a_i(\lambda)$ are strictly decreasing, and the $b_j(\lambda)$ are nondecreasing. These results follow immediately from the previous theorems by considering the functions $\bar{a}_i(\lambda) = a_i(-\lambda)$, which are strictly increasing, and $\bar{b}_j(\lambda) = b_j(-\lambda)$, which are nonincreasing on the interval $(-\Lambda_2, -\Lambda_1)$. For example, the following is a weak version of Theorem A, analogous to Theorem 3.2.

THEOREM 3.2′. *Let $A(\lambda)$ be a matrix of type* (3.1), *and suppose that*

(1) *$a_i(\lambda)$ is strictly decreasing, for $1 \leqq i \leqq n$*;

(2) *$b_j(\lambda)$ is nondecreasing, for $0 \leqq j \leqq n$*;

(3) *$b_j(\lambda)$ has no zeros, for $1 \leqq j \leqq n-1$.*

Let $\lambda' < \lambda''$ be numbers in (Λ_1, Λ_2). Then $\det[A(\lambda)]$ has exactly $c(\lambda') - c(\lambda'')$ different zeros in the interval $(\lambda', \lambda'']$.

Similarly, we can obtain results analogous to Theorems 3.6-3.10 for the matrix (3.8). For example, Theorem 3.10 has the following analogue.

THEOREM 3.10′. *Let $A(\lambda)$ be a matrix of type* (3.8) *in which the functions $b_j(\lambda)$ are nonzero constants for $1 \leqq j \leqq n-1$. Suppose that, for $1 \leqq i \leqq n$,*

(i) *$a_i(\lambda)$ is strictly increasing*;

(ii) $\lim_{\lambda \to \Lambda_1} a_i(\lambda) = -\infty$;

(iii) $\lim_{\lambda \to \Lambda_2} a_i(\lambda) = \infty$.

Then,

(a) *For any* λ_0 *in* (Λ_1, Λ_2), $\det[A(\lambda)]$ *has exactly* $c(\lambda_0)$ *different zeros in the interval* (λ_0, Λ_2);

(b) *The zeros of* $S_i(\lambda)$ *and* $S_{i-1}(\lambda)$ *are interlaced, for* $1 \leqq i \leqq n$;

(c) $\det[A(\lambda)]$ *has exactly* n *different zeros in* (Λ_1, Λ_2).

To illustrate the previous theorem, let $m_1, m_2, \cdots, m_n$ be odd, positive integers, let $\alpha_1, \alpha_2, \cdots, \alpha_n$ be arbitrary real numbers, and let $\beta_1, \beta_2, \cdots, \beta_{n-1}$ be arbitrary, nonzero real numbers. Then the matrix

$$A(\lambda) = \begin{bmatrix} (\lambda^{m_1} - \alpha_1) & \beta_1 & & & & \\ \beta_1 & (\lambda^{m_2} - \alpha_2) & \beta_2 & & & \\ & \beta_2 & (\lambda^{m_3} - \alpha_3) & \beta_3 & & \\ & & \cdots & \cdots & \cdots & \\ & & & \beta_{n-2} & (\lambda^{m_{n-1}} - \alpha_{n-1}) & \beta_{n-1} \\ & & & & \beta_{n-1} & (\lambda^{m_n} - \alpha_n) \end{bmatrix}$$

satisfies Theorem 3.10′. In particular, $p(\lambda) = \det[A(\lambda)]$ has exactly n different real zeros. For example,

$$p(\lambda) = (\lambda^{17} - \alpha_1)(\lambda^{25} - \alpha_2) - \beta^2$$

has exactly two different real zeros, if $\beta \neq 0$, and

$$p(\lambda) = (\lambda^{15} - \alpha_1)(\lambda^{29} - \alpha_2)(\lambda^{53} - \alpha_3) - \beta_2^2(\lambda^{15} - \alpha_1) - \beta_1^2(\lambda^{53} - \alpha_3)$$

has exactly three real zeros, if $\beta_1 \beta_2 \neq 0$.

4. Sturm–Liouville equations. We will apply the results of § 3 to a Sturm–Liouville eigenvalue problem of the form

$$\begin{cases} (p(x, \lambda)u')' + q(x, \lambda)u = 0 \quad \text{for } a \leqq x \leqq b, \\ \alpha_0(\lambda)u(a) + \beta_0(\lambda)u'(a) = 0, \\ \alpha_1(\lambda)u(b) + \beta_1(\lambda)u'(b) = 0, \end{cases} \tag{4.1}$$

where we assume the following:

(i) $p(x, \lambda)$, $(\partial/\partial x)p(x, \lambda)$ and $q(x, \lambda)$ are continuous functions on $[a, b] \times (\Lambda_1, \Lambda_2)$;

(ii) $p(x, \lambda) \geqq k > 0$, for $a \leqq x \leqq b$, $\Lambda_1 < \lambda < \Lambda_2$;

(iii) $\alpha_0(\lambda)$, $\beta_0(\lambda)$, $\alpha_1(\lambda)$, $\beta_1(\lambda)$, are continuous on (Λ_1, Λ_2);

(iv) $\alpha_i(\lambda)^2 + \beta_i(\lambda)^2 \neq 0$, for $i = 0, 1$ and $\Lambda_1 < \lambda < \Lambda_2$;

(v) For $i = 0, 1$, either $\beta_i(\lambda) \equiv 0$ or $\beta_i(\lambda) > 0$, for $\Lambda_1 < \lambda < \Lambda_2$.

We will use the following notation:

$$\begin{cases} p^*(\lambda) = \max\limits_{a \leqq x \leqq b} p(x, \lambda), & p_*(\lambda) = \min\limits_{a \leqq x \leqq b} p(x, \lambda), \\ q^*(\lambda) = \max\limits_{a \leqq x \leqq b} q(x, \lambda), & q_*(\lambda) = \min\limits_{a \leqq x \leqq b} q(x, \lambda). \end{cases} \tag{4.2}$$

The following theorem (with somewhat different notation) can be found in Bôcher [2, Chap. 3] and Ince [5, Chap. 10].

THEOREM D (Sturm Oscillation Theorem). *Suppose that*

(1) *For each* x, $q(x, \lambda)$ *is a strictly increasing function of* λ;

(2) *For each* x, $p(x, \lambda)$ *is a nonincreasing function of* λ;

(3) *If* $\beta_0(\lambda) \not\equiv 0$, *then* $p(a, \lambda)\alpha_0(\lambda)/\beta_0(\lambda)$ *is nondecreasing*;

(4) *If* $\beta_1(\lambda) \not\equiv 0$, *then* $p(b, \lambda)\alpha_1(\lambda)/\beta_1(\lambda)$ *is nonincreasing*;

(5) $\lim_{\lambda \to \Lambda_2} q_*(\lambda)/p^*(\lambda) = \infty$.

Then the eigenvalues of (4.1) *form an infinite, increasing sequence* $\lambda_m < \lambda_{m+1} < \lambda_{m+2} < \cdots$, *which tends to* Λ_2. *The eigenfunction* $\varphi_i(x)$, *corresponding to* λ_i, *has exactly* i *zeros in the interval* (a, b). *Furthermore, suppose that* (4.1) *satisfies either*

(6) $\lim_{\lambda \to \Lambda_1} q^*(\lambda)/p_*(\lambda) = -\infty$, *or*

(7) *There is a number* λ_+ *in* (Λ_1, Λ_2), *so that* $\alpha_0(\lambda_+)\beta_0(\lambda_+) \leqq 0$, $\alpha_1(\lambda_+)\beta_1(\lambda_+) \geqq 0$, *and* $q^*(\lambda_+) \leqq 0$. (*If the coefficient functions in* (4.1) *can be extended continuously to* $\lambda = \Lambda_1$, *we may take* $\lambda_+ = \Lambda_1$.)

Then the sequence of eigenvalues begins with λ_0, *whose eigenfunction* $\varphi_0(x)$ *has no zeros in* (a, b).

We will discretize the boundary value problem (4.1) by finite elements, thereby generating a finite difference scheme. Our goal will then be to show that Sturm sequences may be applied to the finite difference matrix, under the kind of assumptions made in Theorem D.

Recall that the energy inner product for (4.1) is

$$B(u, v) = \frac{\alpha_0(\lambda)}{\beta_0(\lambda)} p(a, \lambda)u(a)v(a) - \frac{\alpha_1(\lambda)}{\beta_1(\lambda)} p(b, \lambda)u(b)v(b) + \int_a^b (-pu'v' + quv)\, dx. \tag{4.3}$$

A weak solution of (4.1) is a function u in the Sobolev space $H^1[a, b]$, such that

$$B(u, v) = 0 \quad \text{for all } v \in H^1[a, b]. \tag{4.4}$$

If (4.4) admits a nontrivial solution $u_0(x)$ for a particular value $\lambda = \lambda_0$, then λ_0 is an eigenvalue, and $u_0(x)$ is a corresponding eigenfunction. (If $\beta_0(\lambda) \equiv 0$, then α_0/β_0 is set equal to 0 in (4.3), and $H^1[a, b]$ is replaced by the subspace of functions $v \in H^1[a, b]$, such that $v(a) = 0$. The case $\beta_1(\lambda) \equiv 0$ is treated similarly. We will carry out the calculations in the generic case $\beta_0(\lambda) \neq 0$, $\beta_1(\lambda) \neq 0$.)

The problem will be discretized using piecewise linear functions, with uniform mesh $h = (b - a)/n$. Consider the partition $x_0 < x_1 < \cdots < x_n$ of the interval $[a, b]$ by the nodes $x_i = a + ih$. The finite element space S_h is the space of continuous functions on $[a, b]$ that are linear on each interval $[x_{i-1}, x_i]$. The inner product (4.3) will now be restricted to S_h, and the integrals in (4.3) will be approximated by quadrature formulas. We will use the midpoint rule for the integral $\int_a^b pu'v'\, dx$, and the trapezoid rule for the integral $\int_a^b quv\, dx$. This defines an inner product $B_h(u, v)$ on S_h. The finite element solution is a function $u \in S_h$, such that

$$B_h(u, v) = 0 \quad \text{for all } v \in S_h. \tag{4.5}$$

A basis $v_0, v_1, \cdots, v_n$ for S_h can be obtained as follows:

$$v_0(x) = \begin{cases} -(x - x_0)/h + 1 & \text{for } x_0 \leqq x \leqq x_1, \\ 0 & \text{elsewhere;} \end{cases} \tag{4.6a}$$

$$v_i(x)=\begin{cases}(x-x_i)/h+1 & \text{for } x_{i-1}\leqq x\leqq x_i,\\ -(x-x_i)/h+1 & \text{for } x_i\leqq x\leqq x_{i+1},\\ 0 & \text{elsewhere,}\end{cases} \tag{4.6b}$$

for $1\leqq i\leqq n-1$;

$$v_n(x)=\begin{cases}(x-x_n)/h+1 & \text{for } x_{n-1}\leqq x\leqq x_n,\\ 0 & \text{elsewhere.}\end{cases} \tag{4.6c}$$

This basis is uniquely determined by the property:

(1) $v_i(x_j)=\delta_{ij}$ $(0\leqq i, j\leqq n)$.

It also satisfies the property:

(2) The support of $v_i(x)$ is contained in the one or two intervals that contain x_i.

A function $u\in S_h$ can be expressed in the form $u(x)=\sum_{i=0}^{n} u_i v_i(x)$, where $u_i=u(x_i)$. Equation (4.5) is now equivalent to the system of equations

$$\sum_{i=0}^{n} u_i B_h(v_i, v_j)=0, \qquad j=0, 1, \cdots, n. \tag{4.7}$$

Because of property (2) above, $B_h(v_i, v_j)=0$ if $|i-j|>1$. A simple calculation shows that

$$\begin{cases} B_h(v_0, v_0)=\dfrac{\alpha_0}{\beta_0}p_0-\dfrac{p_{1/2}}{h}+\dfrac{h}{2}q_0,\\[2ex] B_h(v_n, v_n)=-\dfrac{\alpha_1}{\beta_1}p_n-\dfrac{p_{n-1/2}}{h}+\dfrac{h}{2}q_n,\\[2ex] B_h(v_i, v_i)=-\dfrac{1}{h}(p_{i-1/2}+p_{i+1/2})+hq_i, \qquad 1\leqq i\leqq n-1,\\[2ex] B_h(v_i, v_{i+1})=\dfrac{p_{i+1/2}}{h}, \qquad 0\leqq i\leqq n-1. \end{cases} \tag{4.8}$$

Here, $p_j=p_j(\lambda)$ denotes $p(a+jh, \lambda)$, where j is an integer or half-integer, and similarly for $q_j=q_j(\lambda)$. Multiplying the equations (4.7) by $-h$, we obtain a system of equations:

$$A(\lambda)u=0, \tag{4.9}$$

where $u=(u_0, u_1, \cdots, u_n)^T$ and $A(\lambda)$ is the matrix

$$A(\lambda)=\begin{bmatrix} (b_0-a_0) & -b_0 & & & & & \\ -b_0 & (b_0+b_1-a_1) & -b_1 & & & & \\ & -b_1 & (b_1+b_2-a_2) & & -b_2 & & \\ & & \ddots & \ddots & & \ddots & \\ & & & -b_{n-2} & (b_{n-2}+b_{n-1}-a_{n-1}) & -b_{n-1} & \\ & & & & -b_{n-1} & (b_{n-1}-a_n) \end{bmatrix}. \tag{4.10}$$

The matrix coefficients here are

$$\begin{cases} a_0(\lambda)=\dfrac{h^2}{2}q_0(\lambda)+\dfrac{h\alpha_0(\lambda)}{\beta_0(\lambda)}p_0(\lambda),\\[2ex] a_n(\lambda)=\dfrac{h^2}{2}q_n(\lambda)-\dfrac{h\alpha_1(\lambda)}{\beta_1(\lambda)}p_n(\lambda),\\[2ex] a_i(\lambda)=h^2q_i(\lambda) \quad \text{for } 1\leqq i\leqq n-1,\\[2ex] b_i(\lambda)=p_{i+1/2}(\lambda) \quad \text{for } 0\leqq i\leqq n-1. \end{cases} \tag{4.11}$$

Recall that we have assumed (for $i=0, 1$) that either $\beta_i(\lambda)\equiv 0$, or $\beta_i(\lambda)>0$. If $\beta_0(\lambda)\equiv 0$, then $u_0=0$ and the zeroth row and column in $A(\lambda)$ are omitted. If $\beta_1(\lambda)\equiv 0$, then $u_n=0$ and the nth row and column are omitted. Thus $A(\lambda)$ is an $m\times m$ matrix, where m can be $n-1$, n or $n+1$. We will continue to confine our calculations to the generic case $\beta_i(\lambda)>0$, for $i=0, 1$. In this case, $A(\lambda)$ is an $(n+1)\times(n+1)$ matrix of the form (3.1), where the first row and column are indexed by $i=0$ instead of $i=1$. The term b_0 in (3.1) corresponds to a term b_{-1} in (4.10), and $b_{-1}=b_n=0$ in (4.10).

The assumptions after (4.1) imply that the $a_i(\lambda)$ and $b_j(\lambda)$ are continuous functions, and $b_j(\lambda)>0$. If the assumptions (1)-(4) in Theorem D are satisfied, then the functions $a_i(\lambda)$ are strictly increasing and the $b_j(\lambda)$ are nonincreasing. Therefore we may apply Theorems 3.1, 3.2, and 3.4 to $A(\lambda)$. We will now verify that assumptions (6) or (7) in Theorem D enable us to apply Theorem 3.3, and if we also assume condition (5) in Theorem D, then we may apply Theorem 3.5.

LEMMA 4.1. *Suppose that*

(1) $p(a,\lambda)\alpha_0(\lambda)/\beta_0(\lambda)$ *is a nondecreasing function of* λ;

(2) $p(b,\lambda)\alpha_1(\lambda)/\beta_1(\lambda)$ *is a nonincreasing function of* λ.

Furthermore, assume that either

(3) $\lim_{\lambda\to\Lambda_1} q^*(\lambda)/p_*(\lambda)=-\infty$; *or*

(4) *There is a number* λ_+ *in* (Λ_1,Λ_2), *such that* $\alpha_0(\lambda_+)\beta_0(\lambda_+)\leqq 0$, $\alpha_1(\lambda_+)\beta_1(\lambda_+)\geqq 0$, *and* $q^*(\lambda_+)\leqq 0$.

Then there is a number λ^+ *such that* $a_i(\lambda^+)\leqq 0$, *for* $0\leqq i\leqq n$. (*If* (4) *applies, then* $\lambda^+=\lambda_+$.)

Proof. Suppose that condition (3) applies. Then $\lim_{\lambda\to\Lambda_1} q^*(\lambda)/p_*(\lambda)=-\infty$, and by condition (ii) after (4.1), $p_*(\lambda)\geqq k>0$. This implies that $\lim_{\lambda\to\Lambda_1} q^*(\lambda)=-\infty$. For $1\leqq i\leqq n-1$, $a_i(\lambda)=h^2q_i(\lambda)\leqq h^2q^*(\lambda)$. Therefore $a_i(\lambda)<0$, for λ near Λ_1, $1\leqq i\leqq n-1$. Concerning

$$a_0(\lambda)=\frac{h^2}{2}q_0(\lambda)+\frac{h\alpha_0(\lambda)}{\beta_0(\lambda)}p_0(\lambda),$$

we see condition (1) implies that, for $\Lambda_1<\lambda<\lambda_0$,

$$a_0(\lambda)\leqq\frac{h^2}{2}q_0(\lambda)+\frac{h\alpha_0(\lambda_0)}{\beta_0(\lambda_0)}p_0(\lambda_0)\leqq\frac{h^2}{2}q^*(\lambda)+\frac{h\alpha_0(\lambda_0)}{\beta_0(\lambda_0)}p_0(\lambda_0).$$

Since $\lim_{\lambda\to\Lambda_1} q^*(\lambda)=-\infty$, it follows that $a_0(\lambda)<0$, for λ near Λ_1. Similarly, for

$$a_n(\lambda)=\frac{h^2}{2}q_n(\lambda)-\frac{h\alpha_1(\lambda)}{\beta_1(\lambda)}p_n(\lambda),$$

conditions (2) and (3) imply that $a_n(\lambda)<0$, for λ near Λ_1.

Now suppose that condition (4) applies. Then, for $1\leqq i\leqq n-1$, $a_i(\lambda_+)=h^2q_i(\lambda_+)\leqq h^2q^*(\lambda_+)\leqq 0$. Since $\alpha_0(\lambda_+)\beta_0(\lambda_+)\leqq 0$ and $p_0(\lambda)>0$,

$$a_0(\lambda_+)=\frac{h^2}{2}q_0(\lambda_+)+\frac{h\alpha_0(\lambda_+)}{\beta_0(\lambda_+)}p_0(\lambda_+)\leqq\frac{h^2}{2}q_0(\lambda_+)\leqq\frac{h^2}{2}q^*(\lambda_+)\leqq 0.$$

Similarly,

$$a_n(\lambda_+)=\frac{h^2}{2}q_n(\lambda_+)-\frac{h\alpha_1(\lambda_+)}{\beta_1(\lambda_+)}p_n(\lambda_+)\leqq\frac{h^2}{2}q^*(\lambda_+)\leqq 0.$$

□

LEMMA 4.2. *Assume conditions* (1), (2) *in Lemma* 4.1 *and*

$$\lim_{\lambda \to \Lambda_2} q_*(\lambda)/p^*(\lambda) = \infty. \tag{3}$$

Then $\lim_{\lambda \to \Lambda_2} a_i(\lambda) = \infty$, *for* $0 \leqq i \leqq n$.

Proof. Since $p^*(\lambda) \geqq k > 0$, condition (3) implies that $\lim_{\lambda \to \Lambda_2} q_*(\lambda) = \infty$. For $1 \leqq i \leqq n-1$, $a_i(\lambda) = h^2 q_i(\lambda) \geqq h^2 q_*(\lambda)$. Therefore $\lim_{\lambda \to \Lambda_2} a_i(\lambda) = \infty$, for $1 \leqq i \leqq n-1$. Next, consider

$$a_0(\lambda) = \frac{h^2}{2} q_0(\lambda) + \frac{h\alpha_0(\lambda)}{\beta_0(\lambda)} p_0(\lambda).$$

Because of condition (1),

$$a_0(\lambda) \geqq \frac{h^2}{2} q_0(\lambda) + \frac{h\alpha_0(\lambda_0)}{\beta_0(\lambda_0)} p_0(\lambda_0), \quad \text{for } \lambda_0 < \lambda < \Lambda_2.$$

Therefore $\lim_{\lambda \to \Lambda_2} a_0(\lambda) = \infty$. Similarly, condition (2) implies that $\lim_{\lambda \to \Lambda_2} a_n(\lambda) = \infty$. □

We can now use Theorems 3.1-3.5 to show that Sturm sequences can be applied to the finite difference matrix $A(\lambda)$ in (4.10), for the Sturm-Liouville problem (4.1). We will use our usual notation: $S_i(\lambda)$ is the ith principal minor of $A(\lambda)$; $S_0(\lambda) = 1$; $c(\lambda)$ is the number of sign changes in the sequence $S_0(\lambda), S_1(\lambda), \cdots, S_{n+1}(\lambda)$ after the zero terms have been omitted.

THEOREM 4.1. *Let* $A(\lambda)$ *be the finite difference matrix in* (4.10). *Suppose that conditions* (1)-(4) *in Theorem* D *are satisfied. Then*

(a) *For any numbers* $\lambda' < \lambda''$ *in* (Λ_1, Λ_2), $\det[A(\lambda)]$ *has exactly* $c(\lambda'') - c(\lambda')$ *different zeros in the interval* $[\lambda', \lambda'')$;

(b) *The zeros of* $S_i(\lambda)$ *and* $S_{i-1}(\lambda)$ *are interlaced, for* $1 \leqq i \leqq n+1$.

If (4.1) *also satisfies either* (6) *or* (7) *in Theorem* D, *then*

(c) *For any* λ_0 *in* (Λ_1, Λ_2), $\det[A(\lambda)]$ *has exactly* $c(\lambda_0)$ *different zeros in the interval* (Λ_1, λ_0).

If (4.1) *satisfies conditions* (1)-(5) *and either* (6) *or* (7) *in Theorem* D, *then*

(d) $\det[A(\lambda)]$ *has exactly* $n+1$ *different zeros in* (Λ_1, Λ_2).

Remarks. (1) Theorem 4.1 has been stated for the generic case, where $\beta_0(\lambda) > 0$, $\beta_1(\lambda) > 0$, and $A(\lambda)$ has size $(n+1) \times (n+1)$. It is also valid in case one or both functions $\beta_0(\lambda)$, $\beta_1(\lambda)$ are identically zero. In these cases, $A(\lambda)$ has size $n \times n$ or $(n-1) \times (n-1)$.

(2) Theorem 4.1 allows us to use the bisection method with Sturm sequences to find the kth eigenvalue of the finite difference matrix $A(\lambda)$. This in turn provides a method for approximating the kth eigenvalue of the Sturm-Liouville problem (4.1).

(3) If we interchange the words "increasing" and "decreasing" in Theorem D, we can prove another version of Theorem 4.1 (analogous to Theorem 3.2′ in § 3). This kind of Sturm-Liouville problem occurs in physical applications. For example, the equation

$$\begin{aligned} &\left(\frac{u'}{(\omega - \lambda u^0(x))^2}\right)' + \left(\frac{1}{c(x)^2} - \frac{\lambda^2}{(\omega - \lambda u^0(x))^2}\right) u = 0 \quad \text{for } 0 \leqq x \leqq d, \\ &u(0) = 0 = u'(d) \end{aligned} \tag{4.12}$$

occurs in acoustic problems (see Porter and Reiss [8], [9]).

Acknowledgments. I am grateful to Ivo Babuška for his encouragement, and for many helpful conversations. I would also like to thank the referee, whose comments led to a simplification of Theorem 4.1.

REFERENCES

[1] M. BÔCHER, *The published and unpublished work of Charles Sturm on algebraic and differential equations*, Bull. Amer. Math. Soc., 18 (1912), pp. 1-18.

[2] ———, *Leçons sur les méthodes de Sturm dans la théorie des équations différentielles linéaires*, Gauthier-Villars, Paris, 1917.

[3] W. GIVENS, *A method of computing eigenvalues and eigenvectors suggested by classical results on symmetric matrices*, Appl. Math. Ser. 29, Nat. Bur. Standards (1953), pp. 117-122.

[4] L. GREENBERG AND I. BABUŠKA, *A continuous analogue of Sturm sequences in the context of Sturm-Liouville equations*, SIAM J. Numer. Anal., 26 (1989).

[5] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1956 (reprint).

[6] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[7] C. G. J. JACOBI, *Über eine elementare Transformation eines in Bezug auf jedes von zwei Variablen-Systemen Linearen und Homogenen Ausdrucks*, J. Reine Angew. Math., 53 (1857), pp. 265-270. (Gessammelte Werke, Vol. 3, pp. 583-590).

[8] M. B. PORTER AND E. L. REISS, *A numerical method for acoustic normal modes for shear flows*, J. Sound Vibration, 100 (1985), pp. 91-105.

[9] ———, *A note on the relationship between finite-differences and shooting methods for* ODE *eigenvalue problems*, SIAM J. Numer. Anal., 23 (1986), pp. 1034-1039.

[10] G. SALMON, *Lessons Introductory to the Modern Higher Algebra*, Hodges and Smith, Dublin, 1859. Chelsea, New York, 1964 (reprint).

[11] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[12] C. STURM, *Analyse d'un Mémoire sur la résolution des équations numériques*, Bulletin Universel des Sciences et de l'Industrie, publié sous la direction de M. le Bon. de Ferussac, Paris, Section 1, Vol. 11 (1829), pp. 419-422.

[13] ———, *Extrait d'un Mémoire sur l'intégration d'un système d'équations différentielles linéaires*, Bulletin Universel des Sciences et de l'Industrie, publié sous la direction de M. le Bon, de Ferussac, Paris, Section 1, Vol. 12 (1829), pp. 313-322.

[14] ———, *Mémoire sur la résolution des équations numériques*, Acad. de Sciences Paris, Mémoires des Savants Étrangers 6 (1835), pp. 273-318.

[15] H. WEBER, *Lehrbuch der Algebra*, F. Vieweg, Braunschweig, 1895.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

016

ON THE CONDITIONING OF MULTIPOINT AND INTEGRAL BOUNDARY VALUE PROBLEMS*

F. R. DE HOOG† AND R. M. M. MATTHEIJ‡

Abstract. Linear multipoint boundary value problems are investigated from the point of view of the condition number and properties of the fundamental solution. It is found that when the condition number is not large, the solution space is polychotomic. On the other hand, if the solution space is polychotomic then there exist boundary conditions such that the associated boundary value problem is well conditioned.

Key words. boundary value problem, conditioning, Green function, integral conditions

AMS(MOS) subject classifications. 34B10, 65L10

1. Introduction.

Consider a system of first-order ordinary differential equations

$$\mathscr{L}y := y' - Ay = f, \qquad 0 < t < 1 \tag{1.1}$$

where $A \in L_1^{n\times n}(0, 1)$ and $f \in L_1^n(0, 1)$. We are interested in the solution of (1.1) that satisfies the multipoint boundary condition (BC)

$$\mathscr{B}y := \sum_{i=1}^{N} B_i y(t_i) = b. \tag{1.2}$$

Here, $0 = t_1 < \cdots < t_N = 1$ and the matrices $B_i \in \mathbb{R}^{n\times n}$, $k = 1, \cdots, N$, have been scaled so that, for instance,

$$\sum_{i=1}^{N} B_i B_i^T = I. \tag{1.3}$$

The restriction $t_1 = 0$, $t_N = 1$ has been introduced for notational convenience and is not restrictive provided we allow for the possibility that $B_0 = 0$ and $B_N = 0$.

One of the simplest examples of a multipoint boundary value problem is that of a dynamical system with n states which are observed at different times. Further examples and a description of numerical schemes for the solution of such equations may be found in [12], [1], and [11].

From the theory of boundary value problems, (1.1), (1.2) has a unique solution if $\mathscr{B}Y$ is nonsingular for any fundamental solution Y of $\mathscr{L}$ (see, for example, Keller [8]). In the sequel we assume this is the case. Then, given any fundamental solution Y of (1.1), we may write the solution of (1.1), (1.2) as

$$y(t) = \Phi(t)b + \int_0^1 G(t, s)f(s)\, ds, \qquad 0 \leqq t \leqq 1 \tag{1.4}$$

where

$$\Phi(t) := Y(t)(\mathscr{B}Y)^{-1} \tag{1.5a}$$

* Received by the editors February 27, 1987; accepted for publication (in revised form) May 3, 1988.
† CSIRO, Division of Mathematics and Statistics, P.O. Box 1965, Canberra ACT 2601, Australia.
‡ Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands.

and

$$G(t,s)=\begin{cases}\Phi(t)\sum_{i=1}^{k}B_i\Phi(t_j)\Phi^{-1}(s), & t_k<s<t_{k+1},\quad t>s,\\ -\Phi(t)\sum_{i=k+1}^{N}B_i\Phi(t_j)\Phi^{-1}(s), & t_k<s<t_{k+1},\quad t<s.\end{cases}\tag{1.5b}$$

The function G is the *Green function* associated with (1.1), (1.2).

We can now use (1.4) to examine the conditioning of (1.1), (1.2). Let $|\cdot|$ denote the usual Euclidean norm in $\mathbb{R}^n$ and define

$$\|u\|_\infty := \sup_t |u(t)|, \qquad u\in[L_\infty(0,1)]^n,$$

$$\|u\|_1 = \int_0^1 |u(t)|\,dt, \qquad u\in[L_1(0,1)]^n.$$

Then it follows from (1.3) that

$$\|y\|_\infty \leqq \beta|\mathcal{B}y| + \alpha\|\mathcal{L}y\|_1 \tag{1.6}$$

where

$$\alpha := \sup_{t,s}|G(t,s)| \tag{1.7a}$$

and

$$\beta := \sup_t |\Phi(t)|. \tag{1.7b}$$

The quantities α, β defined by (1.7) serve quite well as condition numbers for the boundary value problem in the sense that they give a measure for the sensitivity of (1.1), (1.2) to changes in the data. Consequently, if α or β is large, we may expect to have difficulties in obtaining an accurate numerical approximation to the solution of (1.1), (1.2).

If α is of moderate size, the solution space of (1.1) has properties that can (and should) be used in the construction of algorithms for calculating an approximate solution of (1.1), (1.2). For the two-point case (i.e., $N=2$), de Hoog and Mattheij [5], [6] have shown that the solution space is dichotomic when α is not too large. A dichotomic solution space (see § 4 for a more detailed discussion of dichotomy) essentially means that nonincreasing modes of the solution space can be controlled by boundary conditions imposed on the left while nondecreasing modes can be controlled by boundary conditions imposed on the right. This concept is the basis for algorithms using decoupling ideas (see, for example, [10], [11]). The aim of this paper is to generalize the results of [5], [6] to (1.1), (1.2) with $N\geqq 2$. In this case the notion of dichotomy has to be generalized, and it turns out that, for well-conditioned problems, the solution space consists of modes that can be controlled at one of the points $t_1,\cdots,t_N$ (see § 4). This has allowed us to generalize the ideas of decoupling to multipoint problems, but that is discussed elsewhere [7].

In general we may say that if $N>n$ there is a redundancy in the number of conditions involved. It is therefore crucial to pick precisely n appropriate points from which modes are actually controlled by suitable conditions. It is quite natural to consider then a limit case of multipoint BC, viz., an integral condition (which incidentally generalizes two and multipoint conditions in an obvious way), so

$$\mathcal{B}y := \int_0^1 B(\tau)y(\tau)\,d\tau = b. \tag{1.8}$$

Such BC arise directly when L_p norms are used to scale the solution (possibly after linearization) as in eigenvalue problems.

We may treat the (discrete) multipoint case separately from (1.8). However, as it turns out, it is possible to construct a general mechanism that handles the integral BC as well. The price to be paid for this is that our proofs will be based on functional analytic arguments and thus are less constructive than could be given for the discrete case. The reward though is that we have been able to get sharp bounds in our estimates, sharpening even the bounds given for the two-point case in [6].

2. Notation and assumptions. In this section we review some basic results that we need later in our analysis. For some general references regarding Green functions we may consult, e.g., [2] and [9].

2.1. Boundary conditions and their normalization. Consider the general *boundary condition* (BC):

$$\mathcal{B}y = b \tag{2.1}$$

where $\mathcal{B}$ is a bounded linear operator from $L^n_{1,1}(0,1)$ to $\mathbb{R}^n$. Note that this includes the BC of type (1.2) and (1.8) as well. By $L^n_{1,1}(0,1)$ we mean those functions the first derivative of which is in $L^n_1(0,1)$. We introduce the norm

$$\|u\|_\infty = \max_{0\leqq t\leqq 1} |u(t)|, \qquad u \in L^n_{1,1}(0,1)$$

where

$$|a| = \left(\sum_{i=1}^{n} a_i^2\right)^{1/2}, \qquad a \in \mathbb{R}^n.$$

For any $a \in \mathbb{R}^n$, $a^T\mathcal{B}$ is a linear functional from $L^n_{1,1}[0,1]$ to $\mathbb{R}$. We define

$$\|a^T\mathcal{B}\|_\infty := \sup_{u\in L^n_{1,1}(0,1)} \frac{|a^T\mathcal{B}u|}{\|u\|_\infty},$$

$$\rho_1(\mathcal{B}) := \max_{a\in\mathbb{R}^n} \frac{\|a^T\mathcal{B}\|_\infty}{|a|} = \|\mathcal{B}\|_\infty,$$

$$\rho_n(\mathcal{B}) := \min_{a\in\mathbb{R}^n} \frac{\|a^T\mathcal{B}\|_\infty}{|a|}.$$

LEMMA 2.1. *Let* $0 < \rho_1(\mathcal{B}) < \infty$. *Then, there exists a matrix* $C \in \mathbb{R}^{n\times n}$ *such that*

$$\|C\mathcal{B}\|_\infty = \rho_1(C\mathcal{B}) = 1$$

and

$$\rho_n(C\mathcal{B}) \geqq \frac{\rho_n(E\mathcal{B})}{\rho_1(E\mathcal{B})} \quad \forall E \in \mathbb{R}^{n\times n}.$$

Proof. If $\rho_n(\mathcal{B}) = 0$, then the result is trivial. We therefore assume $\rho_n(\mathcal{B}) > 0$ and let

$$\mathcal{D} = \{E \in \mathbb{R}^{n\times n} \mid \rho_1(E\mathcal{B}) = 1\}.$$

Since $\rho_n(E\mathcal{B})$ is continuous in E and $\mathcal{D}$ is closed and bounded, it follows that there is a matrix $C \in \mathcal{D}$ such that

$$\rho_n(C\mathcal{B}) \geqq \rho_n(E\mathcal{B}) \quad \forall E \in \mathcal{D}.$$

This is equivalent to the statement of the lemma. □

This now gives us the possibility of scaling the BC, cf. (1.3), in a meaningful way.

Assumption 2.1. In the sequel, we shall assume that the BC (2.1) has been scaled so that

$$\rho_1(\mathcal{B}) = \|\mathcal{B}\|_\infty = 1 \tag{2.2a}$$

and

$$\rho_n(\mathcal{B}) \geqq \rho_n(E\mathcal{B})/\rho_1(E\mathcal{B}) \quad \forall E \in \mathbb{R}^{n\times n}. \tag{2.2b}$$

In addition to Assumption 2.1 we have the following assumption.

Assumption 2.2. Let (1.1), (2.1) have a solution for every $f \in L_1^n(0,1)$ and $b \in \mathbb{R}^n$. Then, $\mathcal{B}Y \in \mathbb{R}^{n\times n}$ is nonsingular, where $Y \in L_{1,1}^{n\times n}(0,1)$ is the solution of

$$\mathcal{L}Y = 0, \qquad Y(0) = F \tag{2.3a}$$

and $F \in \mathbb{R}^{n\times n}$ is nonsingular.

On defining

$$\Phi(t) := Y(t)(\mathcal{B}Y)^{-1}, \tag{2.3b}$$

we can write any function $y \in L_{1,1}^n(0,1)$ as

$$y = \mathcal{P}y + (I-\mathcal{P})y = \mathcal{P}y + \mathcal{G}(\mathcal{L}y), \tag{2.4}$$

where

$$\mathcal{P}y := \Phi(\mathcal{B}y), \tag{2.5a}$$

$$\mathcal{G}f := \int_0^1 \mathcal{G}(t,s)f(s)\,ds, \qquad f \in L_1^n(0,1) \tag{2.5b}$$

and $\mathcal{G}$ is the *Green function* defined by

$$\mathcal{G}(t,s) = \Phi(t)\{H(t,s) - \mathcal{B}(\Phi H(\cdot,s))\}\Phi^{-1}(s) \tag{2.6a}$$

with

$$H(t,s) = \begin{cases} I, & t > s, \\ 0, & t < s \end{cases} \tag{2.6b}$$

(cf. the special case (1.4), where $\mathcal{B}$ is given by (1.2)).

Remark 2.1. The operator $\mathcal{B}$ in the term $\mathcal{B}(\Phi H(\cdot,s))$ above should be interpreted as an extension of $\mathcal{B}$ to an operator from $L_\infty^n(0,1)$ to $\mathbb{R}^n$. Note however that a sensible extension of $\mathcal{B}$ to $L_\infty^n(0,1)$ is assured by the Hahn-Banach theorem.

Remark 2.2. $\mathcal{P}$ is a projection from $L_{1,1}^n(0,1)$ onto the solution space $\{Ya \mid a \in \mathbb{R}^n\}$. Given such a projection $\mathcal{P}$, we can define a linear operator

$$\mathcal{B} = CY^{-1}\mathcal{P}$$

where $C \in \mathbb{R}^{n\times n}$ is a scaling matrix chosen so that (1.1), (2.2a), and (2.2b) hold. Lemma 2.1 ensures the existence of such a matrix.

Remark 2.3. It is easy to verify that the Green function has the form

$$\mathcal{G}(t,s) = \begin{cases} Y(t)(I-E(s))Y^{-1}(s), & t > s, \\ -Y(t)(E(s))Y^{-1}(s), & t < s \end{cases} \tag{2.7}$$

where $E \in L_\infty^{n\times n}(0,1)$. Conversely, given a function of the form (2.7), we have

$$\mathcal{L}\left\{\int_0^1 \mathcal{G}(\cdot,s)f(s)\,ds\right\} = f, \qquad f \in L_1^n(0,1).$$

In addition, if we define

$$(\mathscr{P}y)(t) := y(t) - \int_0^1 \mathscr{G}(t, s)(\mathscr{L}y)(s)\, ds,$$

then

$$(\mathscr{P}y)(t) = y(t) - \int_0^t Y(t)Y^{-1}(s)(\mathscr{L}y)(s)\, ds + Y(t)\int_0^1 E(s)Y^{-1}(s)(\mathscr{L}y)(s)\, ds$$

$$= Y(t)\left\{ Y^{-1}(0)y(0) + \int_0^1 E(s)Y^{-1}(s)(\mathscr{L}y)(s)\, ds \right\}.$$

We can easily verify that $\mathscr{P}$ is a projection. Thus, $\mathscr{B}$ defined by

$$\mathscr{B}y := C\left\{ Y^{-1}(0)y(0) + \int_0^1 E(s)Y^{-1}(s)(\mathscr{L}y(s))\, ds \right\},$$

where $C \in \mathbb{R}^{n\times n}$ is a scaling matrix chosen so that (2.2a), (2.2b) holds, gives a bounded linear operator for which $\mathscr{G}$ is the associated Green function.

2.2. Auerbach's lemma. Let $\mathscr{V}$ be a normed linear space of dimension k with norm denoted by $\|\cdot\|$ and let $\mathscr{V}^*$ be the space of all linear functionals from $\mathscr{V} \to \mathbb{R}$.

Define a norm on $\mathscr{V}^*$ by

$$\|y^*\|^* = \sup_{x\in\mathscr{V}} \frac{y^*(x)}{\|x\|}, \qquad y^* \in \mathscr{V}^*. \tag{2.8}$$

DEFINITION 2.1. A *boundary* of $\mathscr{V}$ is any set

$$\mathscr{D} \subseteq \{y^* \in \mathscr{V}^* \mid \|y^*\|^* \leqq 1\}$$

such that

$$\|x\| = \sup_{y^*\in\mathscr{D}} y^*(x) \quad \forall x \in \mathscr{V}.$$

LEMMA 2.2 (*for Auerbach's lemma see* [4, *Lemma* 4]). *If $\mathscr{D}$ is a closed boundary of $\mathscr{V}$ then there exist $y_i^* \in \mathscr{D}$, $y_j \in \mathscr{V}$; $i, j = 1, \cdots, k$ such that*

$$y_i^*(y_j) = \delta_{ij}, \qquad \|y_i^*\|^* = 1, \qquad \|y_j\| = 1, \qquad i, j = 1, \cdots, k.$$

Since $\{y^* \in \mathscr{V}^* \mid \|y^*\|^* \leqq 1\}$ is a closed boundary, Corollary 2.1 follows immediately.

COROLLARY 2.1. *There exist $y_i^* \in \mathscr{V}^*$, $y_j \in \mathscr{V}$; $i, j = 1, \cdots, k$ such that*

$$y_i^*(y_j) = \delta_{ij}, \qquad \|y_i^*\|^* = 1, \qquad \|y_j\| = 1, \qquad i, j = 1, \cdots, k.$$

3. Conditioning of differential equations. In this section we consider the relation between α and β and the effect of the normalization of the BC as in Assumption 2.1. Recall that for $y \in L^n_{1,1}(0, 1)$ (cf. (2.4))

$$y(t) = \Phi(t)\mathscr{B}y + \int_0^1 \mathscr{G}(t, s)(\mathscr{L}y)(s)\, ds.$$

Hence, on taking norms

$$\|y\|_\infty \leqq \beta|\mathscr{B}y| + \alpha\|\mathscr{L}y\|_1$$

where

$$\beta = \|\Phi\|_\infty = \max_{a\in\mathbb{R}^n} \frac{\|\Phi a\|_\infty}{|a|}, \qquad \alpha = \sup_{t,s} |G(t, s)|.$$

In addition to α and β, it is also useful to consider

$$\mathscr{P}\cdot := Y(\mathscr{B}Y)^{-1}\mathscr{B}\cdot.$$

LEMMA 3.1. $\rho_n(\mathscr{B})\beta \leqq \|\mathscr{P}\|_\infty \leqq \rho_1(\mathscr{B})\beta$.

Proof. The result follows immediately from the definition of $\rho_1(\mathscr{B})$ and $\rho_n(\mathscr{B})$. □

LEMMA 3.2. *Let $\hat{\mathscr{B}}$ be a linear operator from $L^n_{1,1}(0,1)$ to $\mathbb{R}^n$, and let $\hat{\alpha}$ be the constant associated with $\hat{\mathscr{B}}$ and the differential equation* (1.1). *Then,*

$$\hat{\alpha} \leqq (1+\|\hat{\mathscr{P}}\|_\infty)\alpha, \quad \textit{where } \hat{\mathscr{P}}Y = Y(\hat{\mathscr{B}}Y)^{-1}\hat{\mathscr{B}}Y.$$

Proof. Let

$$\hat{\Phi} := Y(\hat{\mathscr{B}}Y)^{-1} \quad \text{and} \quad \hat{\mathscr{G}}f := \int_0^1 \hat{\mathscr{G}}(\cdot, s)f(s)\,ds,$$

where $\hat{\mathscr{G}}$ is defined similarly to $\mathscr{G}$ in (2.6a), i.e., $\mathscr{B}$ replaced by $\hat{\mathscr{B}}$. Clearly, $\hat{\Phi} = Y(\hat{\mathscr{B}}Y)^{-1}$ and consequently $\mathscr{P} = \mathscr{P}\hat{\mathscr{P}}$. That is, $\hat{\mathscr{G}}f = (I-\hat{\mathscr{P}})\mathscr{G}f$, and hence

$$\|\hat{\mathscr{G}}f\|_\infty \leqq (1+\|\hat{\mathscr{P}}\|_\infty)\|\mathscr{G}f\|_\infty.$$

Thus, $\hat{\alpha} \leqq (1+\|\hat{\mathscr{P}}\|_\infty)\alpha$. □

It is clear that the result of Lemmas 3.1 and 3.2 can be combined to give

$$\hat{\alpha} \leqq (1+\rho_1(\tilde{\mathscr{B}})\tilde{\beta})\alpha.$$

Since it has been assumed that (2.2a), (2.2b) hold, we obtain the estimate

$$\hat{\alpha} \leqq (1+\hat{\beta})\alpha. \tag{3.1}$$

Note, however, that α and $\|\mathscr{P}\|_\infty$ are independent of the scaling (2.2a), (2.2b) but that $\rho_1(\mathscr{B})$, $\rho_n(\mathscr{B})$, and β are not. Therefore we examine some of the ramifications of Assumption 2.1.

LEMMA 3.3. $\rho_n(\mathscr{B}) \geqq n^{-1}$.

Proof. Let

$$\mathscr{V} = \{a^T\mathscr{B} \mid a \in \mathbb{R}^n\}.$$

That is, $\mathscr{V}$ are the linear functionals of the form $a^T\mathscr{B}$. Since $\mathscr{B}\Phi = I$, $\dim(\mathscr{V}) = n$. For $\ell \in \mathscr{V}$, define

$$\|\ell\| = \sup_{y \in L^n_{1,1}(0,1)} \frac{(\ell y)}{\|y\|_\infty} = \|\ell\|_\infty.$$

$\mathscr{V}$ equipped with the norm $\|\cdot\|$ is an n-dimensional normed space. From Auerbach's theorem (Corollary 2.1), there exist $\ell_j^* \in \mathscr{V}^*$, $\ell_i \in \mathscr{V}$; $i,j = 1, \cdots, n$ such that

$$\ell_j^*(\ell_i) = \delta_{ij}, \quad \|\ell_j^*\|^* = \|\ell_i\| = 1, \quad i,j = 1, \cdots, n.$$

Clearly, for some $E \in \mathbb{R}^{n\times n}$,

$$a^TE\mathscr{B} = \sum_{i=1}^n a_i\ell_i \quad \forall a = (a_1, \cdots, a_n)^T \in \mathbb{R}^n.$$

Furthermore,

$$\begin{aligned}\|a^TE\mathscr{B}\|_\infty &= \left\|\sum_{i=1}^n a_i\ell_i\right\|_\infty \\ &\geqq \frac{|\sum_{j=1}^n a_j l_j^*(\sum_{i=1}^n a_i\ell_i)|}{\|\sum_{j=1}^n a_j\ell_j^*\|^*} \\ &\geqq \frac{|a|}{\sqrt{n}}.\end{aligned}$$

Thus, $\rho_n(E\mathcal{B}) \geqq 1/\sqrt{n}$. In addition,

$$\|a^T E\mathcal{B}\|_\infty = \left\| \sum_{i=1}^{n} a_i \ell_i \right\|_\infty$$

$$\leqq \sum_{i=1}^{n} |a_i| \, \|\ell_i\|_\infty \leqq n^{1/2}|a|.$$

Thus, $\rho_1(E\mathcal{B}) \leqq n^{1/2}$, and hence from (2.2b)

$$\rho_n(\mathcal{B}) \geqq \frac{\rho_n(E\mathcal{B})}{\rho_1(E\mathcal{B})} \geqq n^{-1}.$$

For boundary conditions of the form (1.2) we can obtain somewhat sharper estimates.

LEMMA 3.4. *For $\mathcal{B}$ given by* (1.2) *and satisfying* (1.1), (2.1), *we have*

$$\rho_n(\mathcal{B}) \geqq N_1^{-1/2}$$

where N_1 is the number of nontrivial matrices B_i in (1.2).

Proof. Without loss of generality, we take $N_1 = N$

$$\|a^T E\mathcal{B}\|_\infty = \sum_{i=1}^{N} |B_i^T E^T a|$$

$$\leqq N^{1/2} \left(a^T E \sum_{i=1}^{N} B_i B_i^T E^T a \right)^{1/2}$$

$$\leqq N^{1/2} \left| E \sum_{i=1}^{N} B_i B_i^T E^T \right|^{1/2} |a|.$$

Thus, $\rho_1(E\mathcal{B}) \leqq N^{1/2} \left| E \sum_{i=1}^{N} B_i B_i^T E^T \right|^{1/2}$. On the other hand,

$$\|a^T E\mathcal{B}\|_\infty = \sum_{i=1}^{N} |B_i^T E^T a|$$

$$\geqq \left(a^T E \sum_{i=1}^{N} B_i B_i^T E^T a \right)^{1/2} \geqq |a| \Big/ \left| \left(E \sum_{i=1}^{N} B_i B_i^T E^T \right)^{-1} \right|^{1/2}.$$

Thus, $\rho_n(E\mathcal{B}) \geqq 1/\left| E \sum_{i=1}^{N} B_i B_i^T E^T \right)^{-1} \Big|^{1/2}$. Now if we take $E = \left(\sum_{i=1}^{N} B_i B_i^T \right)^{-1/2}$, then, from (2.2b), $\rho_n(\mathcal{B}) \geqq \rho_n(E\mathcal{B})/\rho_1(E\mathcal{B}) \geqq N^{-1/2}$. □

For an important class of boundary conditions, the bound in Lemma 3.4 is attained.

LEMMA 3.5. *Let $\mathcal{B}$ be given by* (1.2),

$$\sum_{i=1}^{N} \operatorname{rank}(B_i) = n$$

and N_1 be the number of nontrivial matrices B_i in (1.2). *Then,*

$$\frac{\rho_n(\mathcal{B})}{\rho_1(\mathcal{B})} \leqq N_1^{-1/2}.$$

In addition, (2.2*a*), (2.2*b*) *hold if and only if*

$$\sum_{i=1}^{N} B_i B_i^T = N_1^{-1} I.$$

Proof. Let us assume without loss of generality that $N_1 = N$,

$$B_i^T B_i \eta_i = \sigma_i^2 \eta_i, \quad |\eta_i| = 1, \quad i = 1, \cdots, N$$

and

$$w_1 = 1, \quad w_k = \operatorname{sign}\left\{\eta_k^T B_k^T \sum_{i=1}^{k-1} w_i B_i \eta_i\right\}, \quad k = 2, \cdots, N.$$

Now,

$$\rho_1(\mathscr{B}) = \max_a \left\{\frac{\sum_{i=1}^N |a^T B_i|}{|a|}\right\} \geqq \max_a \left\{\frac{\left|\sum_{i=1}^N w_i a^T B_i \eta_i\right|}{|a|}\right\}$$
$$= \left|\sum_{i=1}^N w_i B_i \eta_i\right| \geqq \left(\sum_{i=1}^N \eta_i^T B_i^T B_i \eta_i\right)^{1/2} = \left(\sum_{i=1}^N \sigma_i^2\right)^{1/2}.$$

This result holds for all singular values σ_i, and we may therefore take $\sigma_i = |B_i|$. Then $\rho_1(\mathscr{B}) \geqq (\sum_{i=1}^N |B_i|^2)^{1/2}$.

In addition, for $\sigma_k \neq 0$,

$$\rho_n(\mathscr{B}) = \min_a \left\{\frac{\sum_{i=1}^N |a^T B_i|}{|a|}\right\} = \sigma_k \min_a \left\{\frac{\sum_{i=1}^N |a^T B_i|}{|a||B_k \eta_k|}\right\}$$
$$\leqq \sigma_k \min_a \left\{\frac{\sum_{i=1}^N |a^T B_i|}{|a^T B_k \eta_k|}\right\} = \sigma_k.$$

Note that the last equality is not valid if $\sum \operatorname{rank}(B_i) > n$. Nor is it valid for an arbitrary vector η_k. Thus,

$$\frac{\rho_n(\mathscr{B})}{\rho_1(\mathscr{B})} \leqq \min_k \frac{\sigma_k}{(\sum_{i=1}^N |B_i|^2)^{1/2}} \leqq N^{-1/2},$$

which proves the first part of the lemma.

Now let (2.2a), (2.2b) hold. From Lemma 3.4 and the result above

$$\sigma_k = N^{-1/2}\left(\sum_{i=1}^N |B_i|^2\right)^{1/2}.$$

Since, σ_k is an arbitrary singular value, all the singular values are equal, and using (2.2a) we obtain that $\sum_{i=1}^N B_i B_i^T = N^{-1} I$.

Finally, let $\sum_{i=1}^N B_i B_i^T = N^{-1} I$. Then, as previously,

$$\rho_1(\mathscr{B}) \geqq \left(\sum_{i=1}^N |B_i|^2\right)^{1/2} = 1 \quad \text{and} \quad \rho_1(\mathscr{B}) \leqq N^{1/2} \left|\sum_{i=1}^N B_i B_i^T\right|^{1/2} = 1.$$

Thus, $\rho_1(\mathscr{B}) = 1$. In addition, as in Lemma 3.4,

$$\rho_n(\mathscr{B}) \geqq 1 \Big/ \left|\left(\sum_{i=1}^N B_i B_i^T\right)^{-1}\right|^{1/2} = N^{-1/2}$$

and since this is the best possible, (2.2b) holds. □

We now have the tools to assess the condition numbers α, β. Let us consider in particular (1.1) and the multipoint BC (1.2),

$$\mathscr{B} y = \sum_{i=1}^N B_i y(t_i),$$

for which we have the following useful properties:

$$\Phi(t) B_i = G^+(t, t_i) - G^-(t, t_i), \qquad i = 1, \cdots, N, \tag{3.2}$$

where

$$G^+(t, t_i) = \lim_{s \to t_i^+} G(t, s), \qquad i = 1, \cdots, N-1, \tag{3.3a}$$

$$G^-(t, t_i) = \lim_{s \to t_i^-} G(t, s), \qquad i = 2, \cdots, N, \tag{3.3b}$$

$$G^+(t, 1) = G^-(t, 0) = 0. \tag{3.3c}$$

THEOREM 3.1. *For $\mathcal{B}$ given by* (2.1) *and satisfying* (2.2a), (2.2b), *we have*

$$\beta \leqq \frac{2N_1\alpha}{\rho_n(\mathcal{B})} \leqq 2N_1\alpha \min(n, N^{1/2}),$$

where N_1 is the number of nontrivial matrices B_1 in (3.2). *If, in addition* $\sum_{i=1}^{N} \operatorname{rank}(B_i) = n$, *then* $\beta \leqq 2N_1\alpha$.

Proof. Without loss of generality, we take $N_1 = N$. From (3.2), (3.3)

$$|\Phi(t)B_i| \leqq 2\alpha,$$

and hence

$$|\Phi(t)| \leqq \left(\sum_{i=1}^{N} |\Phi(t)B_i|^2\right)^{1/2} \left|\left(\sum_{i=1}^{N} B_iB_i^T\right)^{-1}\right|^{1/2}$$

$$\leqq 2\alpha N^{1/2} \left|\left(\sum_{i=1}^{N} B_iB_i^T\right)^{-1}\right|^{1/2}.$$

The first result now follows from the inequality

$$\rho_n(\mathcal{B}) \leqq N^{1/2} \Big/ \left|\left(\sum_{i=1}^{N} B_iB_i^T\right)^{-1}\right|^{1/2}$$

and Lemmas 3.3 and 3.4.

However, if $\sum_{i=1}^{N} \operatorname{rank}(B_i) = n$, it follows from Lemma 3.5 that $|(\sum_{i=1}^{N} B_iB_i^T)^{-1}|^{1/2} = GN^{1/2}$ and this establishes the second part of the theorem. □

Thus, when $\mathcal{B}$ is given by (2.1) and N is not too large, the single parameter α is a suitable measure of the conditioning of the problem. However, as $N \to \infty$ we cannot bound β in terms of α using the results of Theorem 3.1, which suggests that in general it is not possible to obtain such bounds. This is confirmed by the following example.

Example 3.1. Consider the problem

$$\mathcal{L}y = y' + ay, \qquad a > 0.$$

$$\mathcal{B}y = \int_0^1 y(s)\, ds,$$

for which $\alpha = 1$, $\beta = a(1 - e^{-a})$ and $\rho_1(\mathcal{B}) = 1$. Clearly, β becomes unbounded as $a \to \infty$.

Thus, *in general both α and β need to be addressed in a discussion of stability.*

4. Polychotomy. For two-point boundary value problems (i.e., $N = 2$) it has become almost traditional to assume that the solution space

$$\mathcal{S}(t) = \{\Phi(t)c \mid c \in \mathbb{R}^n\}$$

can be separated into a space

$$\mathcal{I}(t) = \{\Phi(t)Pc \mid c \in \mathbb{R}^n\}, \qquad P^2 = P$$

of nondecreasing solutions and a space

$$\mathcal{D}(t) = \{\Phi(t)(I - P)c \mid c \in \mathbb{R}^n\}$$

of nonincreasing solutions. In addition, if neither $\mathcal{S}(t)$ nor $\mathcal{D}(t)$ is trivial (i.e., $P \neq 0, I$), it is usually assumed that the angle $0 < \eta(t) < \pi/2$ between $\mathcal{S}(t)$ and $\mathcal{D}(t)$, defined by

$$\cos \eta(t) = \max_{y_1 \in \mathcal{S}(t), y_2 \in \mathcal{D}(t)} \frac{|y_1^T y_2|}{|y_1||y_2|}$$

is not too small. This has led to the following definition.

DEFINITION 4.1. The solution space is *dichotomic* if there exists a projector P and a constant κ such that

$$|\Phi(t)P\Phi^{-1}(s)| < \kappa, \qquad t > s, \tag{4.1a}$$

$$|\Phi(t)(I-P)\Phi^{-1}(s)| < \kappa, \qquad t < s; \tag{4.1b}$$

κ is called the *dichotomy constant.*

Although a projector always exists such that (4.1) is valid for some constant κ, we are primarily interested in the case when κ is of moderate size. In fact a more precise definition would involve the size of κ as well; we do not dwell on this, however. It turns out that dichotomy is intimately connected with the conditioning of two-point boundary value problems. Specifically, de Hoog and Mattheij [5], [6] have shown the following.

THEOREM 4.1. *When $N = 2$, there exists a projector P such that* (4.1) *holds with $\kappa = \alpha + 4\alpha^2$. Alternatively, if* (4.1) *holds, then there exist matrices $B_1, B_2 \in \mathbb{R}^{n \times n}$ such that $\alpha \leqq \kappa$.*

Thus, if $N = 2$ and α is of moderate size, the solution space is dichotomic (i.e., κ is also of moderate size). Conversely, if the solution space is dichotomic, there is a two-point boundary value problem for which the condition number is not too large.

However, a well-conditioned multipoint problem does not necessarily have a dichotomic solution space as can be seen from Example 4.1.

Example 4.1. Consider the problem

$$y' + 2\lambda(t - \tfrac{1}{2})y = f, \qquad \lambda > 0,$$
$$y(\tfrac{1}{2}) = 1.$$

For this example,

$$\Phi(t) = \exp(-\lambda(t-\tfrac{1}{2})^2),$$
$$y(t) = \Phi(t) + \int_{1/2}^{t} \Phi(t)\Phi^{-1}(s)f(s)\,ds,$$

and hence

$$\alpha = 1 \quad (\text{for all } \lambda).$$

Thus the problem is well conditioned but the fundamental solution now increases on the interval $0 < t < \frac{1}{2}$ and decreases on $\frac{1}{2} < t < 1$. Such behavior is quite common in multipoint problems. Indeed, the results of de Hoog and Mattheij [5], [6] can be used to show that there exist projectors $\hat{P}_i$, $i = 1, \cdots, N-1$ such that

$$|\Phi(t)\hat{P}_i\Phi^{-1}(s)| < \kappa, \qquad t_i < s < t < t_{i+1},$$
$$|\Phi(t)(I-\hat{P}_i)\Phi^{-1}(s)| < \kappa, \qquad t_i < t < s < t_{i+1},$$

where κ is of moderate size if α is not large. Thus, on each interval $t_i < t < t_{i+1}$, $i = 1, \cdots, N-1$ the solution space is dichotomic.

However, the examination of a number of well-conditioned multipoint problems has suggested that additional structure is present in the solution space. This leads to the following generalization of dichotomy.

DEFINITION 4.2. The solution space $\mathcal{S}(t)$ is *polychotomic* if, for some $M \in \mathbb{N}$, and $0 = x_1 \leqq x_2 \leqq \cdots \leqq x_M = 1$, there exist projectors P_k, $k = 1, \cdots, M$ and a constant κ such that

$$\sum_{k=1}^{M} P_k = I, \qquad P_i P_j = P_j P_i = \delta_{ij} P_j,$$

$$\left| \Phi(t) \sum_{j=1}^{k} P_j \Phi^{-1}(s) \right| < \kappa, \quad x_k < s < x_{k+1}, \quad t > s, \tag{4.2a}$$

$$\left| \Phi(t) \sum_{j=k+1}^{M} P_j \Phi^{-1}(s) \right| < \kappa, \quad x_k < s < x_{k+1}, \quad t < s. \tag{4.2b}$$

In § 5 we show that the concept of polychotomy is closely related to the conditioning of multipoint boundary value problems in the sense that κ will be of moderate size when α is not too large. It turns out that this relationship can be exploited in the construction of efficient numerical schemes for the solution of (1.1), (1.2); this is discussed in detail in [7].

5. Bounds for polychotomy. In this section we show how the condition number α can be used to obtain bounds for κ. Initially we consider separable boundary conditions.

5.1. Separable boundary conditions.

DEFINITION 5.1. The boundary condition (1.2) is called *separable* if

$$\sum_{i=1}^{N} \operatorname{rank}(B_i) = n.$$

Thus for separable boundary conditions, the solution space consists of a number of modes each of which is controlled by a condition at one of the points when $\operatorname{rank}(B_i) \neq 0$.

We shall see that when the boundary condition (1.2) is separable, the solution space is polychotomic with constant $\kappa = \alpha$. Before we can show this, however, some preliminary results are required.

LEMMA 5.1. *If* $C_k \in \mathbb{R}^{n \times n}$, $k = 1, \cdots, N$

$$\sum_{k=1}^{N} C_k = I \quad \textit{and} \quad \sum_{k=1}^{N} \operatorname{rank}(C_k) = n,$$

then C_k, $k = 1, \cdots, N$ *are projectors* (*i.e.*, $C_i C_j = C_j C_i = \delta_{ij} C_j$).

Proof. The result follows from the arguments used in [6, Thm. 3.2]. □

LEMMA 5.2. *For* $E_k \in \mathbb{R}^{n \times n}$, $k = 1, \cdots, N$, *let*

$$\sum_{k=1}^{N} E_k = I, \qquad \sum_{k=1}^{N} \operatorname{rank}(E_k) = n,$$

and define

$$\hat{G}(t, s) = \begin{cases} Y(t) \sum_{k=1}^{i} E_k Y^{-1}(s), & t_i < s < t_{i+1}, \quad t > s, \\ -Y(t) \sum_{k=i+1}^{N} E_k Y^{-1}(s), & t_i < s < t_{i+1}, \quad t < s, \end{cases}$$

where Y *is a fundamental solution of* (1.1). *Then there exists a boundary condition*

$$\hat{\mathcal{B}} y := \sum_{i=1}^{N} \hat{B}_i y(t_i) \tag{5.1}$$

satisfying $\operatorname{rank}(\hat{B}_i) = \operatorname{rank}(E_i)$ *and*

$$\sum_{i=1}^{N} \hat{B}_i \hat{B}_i^T = N_1^{-1} I$$

such that $\hat{G}$ *is the Green function associated with* (1.1), (5.1) *and* N_1 *is the number of nontrivial matrices* E_i.

Proof. Consider the LQ^T decomposition

$$[E_1 Y^{-1}(t_1) | E_2 Y^{-1}(t_2)| \cdots | E_N Y^{-1}(t_N)] = LQ^T$$

where $L \in \mathbb{R}^{n\times n}$ is lower triangular and nonsingular and $Q \in \mathbb{R}^{(N+1)n\times n}$ is orthogonal (i.e., $Q^TQ = I$). Now define $\hat{B}_i \in \mathbb{R}^{n\times n}$, $k = 1, \cdots, N$ by

$$[\hat{B}_1 | \hat{B}_2| \cdots | \hat{B}_N] := N_1^{-1} Q^T.$$

If we define

$$\hat{\Phi}(t) := Y(t)(\hat{\mathcal{B}} Y)^{-1},$$

we see that $\hat{\Phi}(t) = Y(t)L$. Then it is easy to verify that $\hat{G}$ is the Green function associated with (1.1), (5.1), viz.,

$$\hat{G}(t, s) = \begin{cases} \hat{\Phi}(t) \sum_{i=1}^{K} \hat{B}_i \Phi(t_i)\Phi^{-1}(s), & t > s, \\ -\hat{\Phi}(t) \sum_{i=k+1}^{N} \hat{B}_i \Phi(t_i)\Phi^{-1}(s), & t < s \end{cases}$$

can be identified with $\hat{G}(t, s)$. □

The relationship between polychotomy and the condition number for separable boundary conditions is now straightforward. Specifically we have the following theorem.

THEOREM 5.1. *If the boundary condition* (1.2), *is separable, then the solution space is polychotomic with* $\kappa \leqq \alpha$.

Conversely, if the solution space of (1.1) *is polychotomic with constant* κ, *then there exists a separable boundary condition* (1.2), *satisfying Assumption* 2.1, *such that* $\alpha \leqq \kappa$.

Proof. If the boundary condition (1.2) is separable

$$\sum_{i=1}^{N} \operatorname{rank}(B_i) = n$$

and

$$\sum_{i=1}^{N} B_i \Phi(t_i) = I \quad \text{(cf. (2.3b))}.$$

Thus

$$\sum_{i=1}^{N} \operatorname{rank}(B_i \Phi(t_i)) = n$$

and from Lemma 5.1,

$$P_i = B_i \Phi(t_i), \qquad i = 1, \cdots, N$$

are projectors. On substituting for P_i in the Green function (1.5) and comparing the resulting expression with the definition of polychotomy (see Definition 5.1), we find that (4.2) holds with $\kappa = \alpha$, $M = N$ and $x_j = t_j$.

If on the other hand the solution is polychotomic, then

$$|G(t, s)| \leqq \kappa$$

where

$$G(t,s) = \begin{cases} Y(t) \sum_{i=1}^{k} P_i Y^{-1}(s), & x_k < s < x_{k+1}, \quad t > s, \\ -Y(t) \sum_{i=k+1}^{M} P_i Y^{-1}(s), & x_k < s < x_{k+1}, \quad t < s \end{cases}$$

and

$$\sum_{i=1}^{M} P_i = I, \qquad P_i P_j = P_j P_i = \delta_{ij} P_j.$$

But from Lemmas 5.2 and 3.5 there exists a separable boundary condition of the form (1.2) which satisfies Assumption 2.1 and is such that G is the Green function associated with (1.1), (1.2) when $N = M$ and $t_i = x_i$. □

5.2. General boundary condition. We again turn to the general BC (2.1) and show how we can select appropriate separable BC from them; this is based on the theory given in § 2.

Let

$$\mathcal{S} = \{ Ya \mid a \in \mathbb{R}^n \}$$

with

$$\|y\| = \|y\|_\infty, \qquad y \in \mathcal{S}.$$

Clearly, $\mathcal{S}$ equipped with the norm $\|\cdot\|$ is a normed space of dimension n. In addition,

$$\mathcal{D} = \{ y^* \in \mathcal{S}^* \mid y^*(y) = c^T y(t), \quad |c| = 1, \quad 0 \leqq t \leqq 1 \}$$

is a closed boundary for $\mathcal{S}$. Hence, from Auerbach's lemma (Lemma 2.2) there exist $y_j^* \in \mathcal{D}$, $y_i \in \mathcal{S}$; $i, j = 1, \cdots, n$ such that

$$y_j^*(y_i) = \delta_{ij}, \quad \|y_j^*\|^* = 1, \quad \|y_i\|_\infty = 1, \quad i, j = 1, \cdots, n.$$

That is, there exist $c_j \in \mathbb{R}_1^n$, $|c_j| = 1$, points t_j with $0 \leqq t_j \leqq 1$, $j = 1, \cdots, n$ and $y_i \in \mathcal{S}$, $i = 1, \cdots, n$ such that

$$c_j^T y_i(t_j) = \delta_{ij}, \quad |c_j| = \|y_i\|_\infty = 1, \quad i, j = 1, \cdots, n. \tag{5.2}$$

Furthermore,

$$c_j = y_j(t_j),$$

and hence

$$c_i^T c_j = 0 \quad \text{if } i \neq j \text{ and } t_i = t_j. \tag{5.3}$$

Let

$$(\hat{\mathcal{P}} y)(t) := \sum_{i=1}^{n} y_i(t) c_i^T y(t_i).$$

Thus,

$$\|\hat{\mathcal{P}} y\|_\infty \leqq \sum_{i=1}^{n} \|y_i\|_\infty \|y\|_\infty$$

$$\leqq n \|y\|_\infty.$$

Hence

$$\|\hat{\mathscr{P}}\|_\infty \leqq n$$

and, as in Lemma 3.2, we find that

$$\begin{aligned}\hat{\alpha} &\leqq (1+\|\hat{\mathscr{P}}\|_\infty)\alpha \\ &\leqq (n+1)\alpha.\end{aligned}$$

In addition, we have

$$\hat{\mathscr{B}}\hat{\Phi} = I$$

where

$$\hat{\Phi} = N_1^{1/2}[y_1|\cdots|y_n], \tag{5.4}$$

$$\begin{aligned}\hat{\mathscr{B}}y &:= \sum_{i=1}^{N} \hat{\mathscr{B}}y(t_i), \\ B_k &= N_1^{-1/2}\begin{bmatrix}0\\ C_k\\ 0\end{bmatrix} \leftarrow k\text{th position},\end{aligned} \tag{5.5}$$

and N_1 is the number of *distinct points* in the set $\{t_k\}$. From (5.2), (5.3)

$$\sum_{k=1}^{n} \hat{B}_k \hat{B}_k^T = N_1^{-1} I,$$

and hence from Lemma 3.5, the boundary condition $\tilde{B}$ defined by (5.5), which is clearly separable, satisfies (2.2a), (2.2b). Finally from (5.2), (5.5)

$$|\hat{\Phi}(t)| \leqq N_1^{1/2} n^{1/2}.$$

Thus, we have shown the following theorem.

THEOREM 5.2. *For a general* BC (2.1) *we can construct a separable* BC $\hat{\mathscr{B}}$ *of the form* $\hat{\mathscr{B}}y := \sum_{i=1}^{n} \hat{B}_i y(t_i)$, *with* $t_i \in [0,1]$, *such that* $\hat{\mathscr{B}}$ *satisfies* (2.2a) *and* (2.2b) *and for which* (cf. (1.7))

$$\hat{\beta} := \sup_t |\hat{\Phi}(t)| \leqq n, \qquad \hat{\alpha} := \sup_{s,t} |\hat{G}(s,t)| \leqq (n+1)\alpha.$$

COROLLARY 5.1. *If the* BVP (1.1), (2.1) *has a condition number* α, *then the solution space is polychotomic with*

$$\kappa \leqq (n+1)\alpha.$$

Note that the result of this corollary is somewhat different from Theorem 3.16 of [6], where bounds are derived $\sim \alpha^2$ for the two-point case. For large α we may therefore say that this more general result is sharper, though not constructive.

REFERENCES

[1] R. P. AGARWAL, *The numerical solution of multipoint boundary value problems.* J. Comp. Appl. Math., 5 (1979), pp. 17-24.

[2] F. V. ATKINSON, *Discrete and Continuous Boundary Problems*, Academic Press, New York, 1964.

[3] C. DE BOOR AND H.-O. KREISS, *On the condition of the linear system associated with discretized* BVPs *of* ODEs, SIAM J. Numer. Anal., 23 (1985), pp. 936-939.

[4] E. W. CHENEY AND K. H. PRICE, *Minimal projections in approximation theory*, in Approximation Theory, A. Talbot, ed., Academic Press, New York, 1970, pp. 261-289.

[5] F. R. DE HOOG AND R. M. M. MATTHEIJ, *The role of conditioning in shooting techniques*, in Numerical Boundary Value ODEs, U. Ascher and R. Russell, eds., Birkhäuser, Boston, 1985, pp. 21-54.

[6] ———, *On dichotomy and well conditioning in* BVP, SIAM J. Numer. Anal., 24 (1987), pp. 89-105.

[7] ———, *An algorithm for solving multipoint boundary value problems*, Computing, 38 (1987), pp. 219-234.

[8] H. B. KELLER, *Numerical Solution of Two-Point Boundary Value Problems*, CBMS-NSF Regional Conference Series in Applied Mathematics 24, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1976.

[9] W. S. LOUD, *Generalized inverses and generalized Green functions*, J. Soc. Indust. Appl. Math., 14 (1966), pp. 342-369.

[10] R. M. M. MATTHEIJ, *Decoupling and stability of algorithms for boundary value problems*, SIAM Rev., 27 (1985), pp. 1-44.

[11] R. M. M. MATTHEIJ AND G. W. M. STAARINK, *An efficient algorithm for solving general linear two-point* BVP, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 745-763.

[12] W. WELSH AND T. OJIKA, *Multipoint boundary value problems with discontinuities*, I. *Algorithms and Applications*, J. Comp. Appl. Math., 6 (1980), pp. 133-143.

SIAM J. MATH. ANAL
Vol. 20, No. 1, January 1989

017

THE BOHL TRANSFORMATION AND OSCILLATION OF LINEAR DIFFERENTIAL SYSTEMS*

STUART GOFF† AND DONALD F. ST. MARY‡

Abstract. The Bohl transformation is a nonlinear transformation that, like the Riccati transformation, relates linear and nonlinear differential equations. Recently, we have extended that transformation to differential systems. In this paper the utility of the transformation for extending scalar oscillation results to linear differential systems is demonstrated.

Key words. oscillation, linear differential systems, nonlinear transformation

AMS(MOS) subject classification. 34C10, 34C20

1. Introduction. The use of transformations in studying qualitative properties of linear differential systems has proven most effective. Probably the best-known such transformation is the Riccati transformation, which has been used in numerous ways. Another example is the Prüfer or generalized polar coordinate transformation, which has been employed most successfully in the study of eigenvalue problems and is used in the current development. In [6], [7], the authors were successful in extending the Bohl transformation to systems of differential equations. In this paper we shall demonstrate the use of the Bohl transformation in establishing oscillation criteria for the self-adjoint differential system

$$X''+P(t)X=0. \tag{1}$$

In general the Bohl transformation can be described as follows: for $X_1(t)$ and $X_2(t)$ solutions of (1) on $[a,\infty)$ satisfying the initial conditions $X_1(a)=X_2'(a)=E$ and $X_1'(a)=X_2(a)=0$, E the $n\times n$ identity matrix, set $V(t)\equiv[X_1^2(t)+X_2^2(t)]^{1/2}$, then $V(t)$ is the *Bohl transformation* of the pair $X_1(t)$, $X_2(t)$ and satisfies

$$V''+P(t)V=V^{-3} \tag{2}$$

on $[a,\infty)$. Equation (2) is the *Bohl differential system* corresponding to (1).

In [7] it is established that, under appropriate hypotheses on P, system (2) has a positive definite Hermitian solution on $[a,\infty)$ satisfying $V(a)=E$, $V'(a)=0$. Furthermore, it is shown that if $V(t)$ is a positive definite Hermitian solution of (2) on $[a,\infty)$ satisfying $V(a)=E$, $V'(a)=0$, then $X_1(t)=V(t)C(t;a,V^{-2}(t))$, $X_2(t)=V(t)S(t;a,V^{-2}(t))$ are conjoined solutions of (1) on $[a,\infty)$ whose Wronskian is E where $C(t;a,Q(t))$ and $S(t;a,Q(t))$ are the matrix analogues of the sine and cosine functions, respectively. These matrix trigonometric functions have been developed in the context of studying the generalized polar coordinate transformation. See, e.g., [1], [3], and [12].

The Bohl transformation is established in [7] under the assumption that $P(t)$ is an $n\times n$ functionally commutative Hermitian matrix the elements of which are continuous complex-valued functions of a real variable $t\in[a,\infty)$. To say $P(t)$ is *functionally commutative* on $[a,\infty)$ means that $P(s)P(t)=P(t)P(s)$, for all $s,t\in[a,\infty)$.

* Received by the editors October 1, 1986; accepted for publication (in revised form) April 26, 1988.
† Department of Mathematics, Keene State College, Keene, New Hampshire 03431.
‡ Department of Mathematics, University of Massachusetts, Amherst, Massachusetts 01003. The work of this author was partially supported by National Science Foundation grant PRM-8114310.

For our purposes a *solution* of (1) is an $n \times n$ matrix function, the elements of which are twice continuously differentiable, which satisfies (1) on $[a, \infty)$. For the discussion of the oscillatory behavior of solutions of (1), it is necessary to restrict our attention to the class of solutions called conjoined. A solution X of (1) defined on an interval I is said to be *conjoined* if X is not identically singular on any subinterval of I and X satisfies

$$X^*(t)X'(t) \equiv X'^*(t)X(t)$$

on I. The system (1) is said to be *oscillatory* on $[a, \infty)$ if there exists a conjoined solution $X(t)$ for which there is a sequence $\{t_j\}$, $t_j \to \infty$, such that $\det X(t_j) = 0$, $j = 1, 2, \cdots$. It follows from the Sturm-type separation theorem [9] that if (1) is oscillatory then every conjoined solution is singular on some sequence $\{t_j\}$, $j \to \infty$. Thus, (1) is *nonoscillatory* if there exists a conjoined solution $X(t)$ for which $\det X(t) \neq 0$ for $t > t_0 \geq a$, for some t_0.

2. Oscillation via the Bohl transformation. In [7], with $P(t)$ as described above, we have established an oscillation theorem that is one of the central elements to be used in the further development of the theory. For purposes of convenience we state it as a lemma.

LEMMA 1. *Let $V(t)$ be a positive definite Hermitian solution of* (2) *on* $[a, \infty)$ *for which $V(a) = E$, $V'(a) = 0$. Then* (1) *is oscillatory if and only if* $\int_a^\infty \operatorname{tr}[V^{-2}(t)]\, dt = \infty$.

To continue our discussion of oscillation we further require that $P(t)$ is analytic on $[a, \infty)$. When we say a matrix $M(t)$ is *analytic* on an interval I, we mean that for every $t_0 \in I$, each element of $M(t)$ (and thus $M(t)$ itself) can be represented as a Taylor series centered at t_0 converging in some neighborhood of t_0.

Some properties of analytic Hermitian matrices follow. A proof of Lemma 2 may be found in [14, pp. 36–45].

LEMMA 2. *Let $M(t)$ be an $n \times n$ analytic Hermitian matrix function on I. Then for $i = 1, \cdots, n$ eigenvalues $\lambda_i(t)$ (perhaps repeated), and an orthonormal set of eigenvectors $z_i(t)$ and eigenprojections $G_i(t)$ (perhaps repeated) for $M(t)$ can be chosen such that $\lambda_i(t)$, $z_i(t)$, and $G_i(t)$ are analytic on I.*

LEMMA 3. *If $M(t)$ is an analytic, Hermitian, functionally commutative matrix on I, then $M(t)M'(t) = M'(t)M(t)$ on I.*

Proof. $M(s)M(t) = M(t)M(s)$ and so $M(s)M'(t) = M'(t)M(s)$ for all $s, t \in I$. Thus, when we let $s = t$, the result follows. □

If $\lambda_i(t)$ and $\lambda_j(t)$ are analytic eigenvalues of $M(t)$ on I, then we say that $\lambda_i(t)$ and $\lambda_j(t)$ are *distinct* if $\lambda_i(t) \not\equiv \lambda_j(t)$ on I. Because $\lambda_i(t)$ and $\lambda_j(t)$ are analytic, it is clear that if they are distinct on I, then they must also be distinct on every subinterval of I. We do allow for the possibility that $\lambda_i(t) = \lambda_j(t)$ for some $t \in I$, but this will be true only on a set of isolated points.

The next lemma is an immediate consequence of the Spectral Representation Theorem (see, e.g., [10, p. 175]), the previous two lemmas, and a result in [5, p. 38] which establishes that the eigenprojections are constant.

LEMMA 4. *Let $M(t)$ be an analytic, Hermitian, functionally commutative matrix on I, and let r be the number of distinct analytic eigenvalues of $M(t)$. Then, $M(t) = \sum_{i=1}^{r} \lambda_i(t) G_i$, where the analytic eigenprojections G_i are constant Hermitian matrices for which $G_i^2 = G_i$, $G_i G_j = 0$ (if $i \neq j$) and $\sum_{i=1}^{r} G_i = E$.*

LEMMA 5. *If $M(t)$ is an analytic, Hermitian, functionally commutative matrix on I, then $M(t)$ may be diagonalized by a constant unitary matrix.*

Proof. By Lemma 3, $M(t)$ commutes with its derivative and thus [5, p. 37] each of the eigenspaces of the analytic eigenvalues of $M(t)$ has a constant basis. Hence,

an orthonormal set of constant eigenvectors for $M(t)$ may be found. Let the matrix H be such that its columns are these constant eigenvectors. Then $H^*M(t)H = Q(t)$ is a diagonal matrix consisting of the analytic eigenvalues of $M(t)$. By construction, H is a constant unitary matrix. □

Our first oscillation theorem for systems will use this matrix H along with the transformation described in the following lemma. A proof of this lemma may be found in [15, pp. 393-394].

LEMMA 6. *For $R(t)$ and $P(t)$ continuous Hermitian matrix functions, let $H(t)$ be an $n \times n$ matrix function such that H and RH' are absolutely continuous on $[a, \infty)$ and $(RH')^*H = H^*(RH')$ on $[a, \infty)$. Put $\mathcal{R}(t) = H^*RH$ and $\mathcal{P}(t) = H^*[(RH')' + PH]$. Then $H^*[(RU')' + PU] = (\mathcal{R}X')' + \mathcal{P}X$ where $U = HX$. Furthermore, if H is nonsingular on $[a, \infty)$ then the system $(RU')' + PU = 0$ is oscillatory on $[a, \infty)$ if and only if the system $(\mathcal{R}X')' + \mathcal{P}X = 0$ is oscillatory on $[a, \infty)$.*

THEOREM 7. *Let $P(t)$ be analytic, Hermitian, and functionally commutative on $[a, \infty)$. Let H be a constant unitary matrix which diagonalizes $P(t)$ and put $Q(t) = H^*P(t)H$. Then (1) is oscillatory on $[a, \infty)$ if and only if $X'' + Q(t)X = 0$ is oscillatory on $[a, \infty)$.*

Proof. H exists as described because of Lemma 5. Moreover, H trivially satisfies the hypotheses of Lemma 6. In the notation of that lemma, $R(t) = E$ and thus $\mathcal{R} = H^*H = E$ since H is unitary, and $\mathcal{P} = H^*PH = Q$ since $H' = 0$. The result follows by applying the second conclusion of Lemma 6. □

COROLLARY 8. *Let $P(t)$ be analytic, Hermitian, and functionally commutative on $[a, \infty)$. Then, the system (1) is oscillatory on $[a, \infty)$ if and only if the scalar equation $x'' + \lambda(t)x = 0$ is oscillatory on $[a, \infty)$ for at least one of the analytic eigenvalues $\lambda(t)$ of $P(t)$.*

Proof. Because of Lemma 5, Theorem 7 is applicable. Since $Q(t) = \operatorname{diag}(\lambda_1(t), \cdots, \lambda_n(t))$ where $\{\lambda_i(t) \mid i = 1, \cdots, n\}$ are the (perhaps repeated) analytic eigenvalues of $P(t)$, it is easy to see that the system $X'' + Q(t)X = 0$ is oscillatory on $[a, \infty)$ if and only if $x'' + \lambda(t)x = 0$ is oscillatory on $[a, \infty)$ for at least one of these eigenvalues. The result now follows from Theorem 7. □

In Theorem 2.1 of [2] a similar result is obtained in which $P(t)$ is assumed to be continuous, symmetric, functionally commutative, and implicitly conservative on $[a, \infty)$ (see Theorem 8 and the general *conservative* hypothesis on p. 108 of [4]). It is easy to demonstrate matrices $P(t)$ that are conservative but not analytic and vice versa, while still satisfying the other hypotheses.

We will next concern ourselves with the extension to systems of M. Ráb's so-called "Haupsatz" (sic) [11, p. 339]:

Let $p(t)$ be continuous on $[a, \infty)$. The scalar differential equation $x'' + p(t)x = 0$ is oscillatory on $[a, \infty)$ if and only if there exists a C^1 function $g(t) > 0$ which satisfies

$$\int_a^\infty \exp\left(-2\int_a^x \frac{1}{g^2(s)}\left[\int_a^s [(g'(t))^2 - p(t)g^2(t)]\,dt + k\right]ds\right)dx = \infty$$

for every constant k.

Our integral condition will involve the solution of a certain differential system; in the scalar case, this solution will result in the above exponential function. To this end, define the matrix function

$$Q(t; g, A) = \frac{1}{g^2(t)}\left\{A + \int_a^t [(g'(s))^2 E - g^2(s)P(s)]\,ds\right\}$$

where $g(t) > 0$ is C^1 on $[a, \infty)$ and A is a constant Hermitian matrix. In addition, let

$Y_{g,A}(t; T)$ denote the solution on $[a, \infty)$ of the initial value problem

$$Y' = Q(t; g, A)Y, \qquad Y(T) = E. \tag{3}$$

In the next lemma we list, in the context of (3), some standard properties of linear differential systems we shall need.

LEMMA 9. *For all $s, t, T \in [a, \infty)$ and with $Y_{g,A}$ as defined above*:

(a) $Y_{g,A}(t; s)Y_{g,A}(s; T) = Y_{g,A}(t; T)$,

(b) $Y_{g,A}^{-1}(s; t) = Y_{g,A}(t; s)$,

(c) $d/dt\,[Y_{g,A}(s; t)] = -Y_{g,A}(s; t)Q(t; g, A)$.

Before proceeding with the extension of the Hauptsatz to systems, let us note some additional properties of the solutions of (1) resulting from the analyticity of $P(t)$. The next lemma follows from standard existence theorems (see, e.g., [8, p. 70]).

LEMMA 10. *Let $P(t)$ be analytic on $[a, \infty)$. Then every solution of* (1) *is analytic on $[a, \infty)$.*

In [5, p. 38] it has been shown that a Hermitian analytic matrix function that commutes with its derivative is functionally commutative. In [7], the existence on $[a, \infty)$ of Hermitian conjoined solutions of (1) that satisfy $X'X = XX'$ is established. Thus, in view of these results and the previous lemma we have the following result.

LEMMA 11. *Let $P(t)$ be analytic, Hermitian, and functionally commutative on $[a, \infty)$. Then, the Hermitian conjoined solutions of* (1) *are functionally commutative on $[a, \infty)$.*

LEMMA 12. *If $X(t)$ is a functionally commutative solution of* (1) *on $[a, \infty)$, then $P(s)X(t) = X(t)P(s)$ for all $s, t \in [a, \infty)$.*

Proof. $X(s)X(t) = X(t)X(s)$ leads to $X''(s)X(t) = X(t)X''(s)$, resulting in $P(s)X(s)X(t) = X(t)P(s)X(s)$, and so $P(s)X(t)X(s) = X(t)P(s)X(s)$. Since the singularities of $X(s)$ are isolated, the result follows. □

We are now prepared to extend Ráb's Hauptsatz to systems.

THEOREM 13. *Let $P(t)$ be analytic, Hermitian, and functionally commutative on $[a, \infty)$. Then, the system* (1) *is oscillatory on $[a, \infty)$ if and only if there exists a C^1 function $g(t) > 0$ that satisfies*

$$\int_c^\infty \operatorname{tr}\{[Y_{g,A}(t; \tau)Y_{g,A}^*(t; \tau)]^{-1}\}\, dt = \infty \tag{4}$$

for some $c, \tau \in [a, \infty)$ and for every $n \times n$ constant Hermitian matrix A that is a linear combination of the eigenprojections of $P(t)$.

Proof. Let (4) hold for c, τ, A and $g(t)$ as described and assume that (1) is nonoscillatory on $[a, \infty)$. In view of the remarks preceding Lemma 11 along with Lemmas 10 and 11, there exists a pair $X_1(t)$ and $X_2(t)$ of functionally commutative, analytic Hermitian conjoined solutions of (1) whose Wronskian is E, both of which are nonsingular on some interval $[t_0, \infty)$. Letting $X(t)$ denote either $X_1(t)$ or $X_2(t)$, we have by the Riccati transformation [13, p. 101] that $(X'X^{-1})' + (X'X^{-1})^2 + P = 0$. Multiplying this equation by $g^2(t)$, integrating from t_0 to t followed by an integration by parts, and then completing the square, we obtain

$$g^2(t)X'(t)X^{-1}(t) = g^2(t_0)X'(t_0)X^{-1}(t_0) - \int_{t_0}^t [g(s)X'(s)X^{-1}(s) - g'(s)E]^2\, ds$$

$$+ \int_{t_0}^t [(g'(s))^2 E - g^2(s)P(s)]\, ds.$$

Since X is conjoined, $gX'X^{-1} - g'E$ is Hermitian and so its square must be nonnegative

definite. Thus we have

$$(5)\qquad g^2(t)X'(t)X^{-1}(t) \geqq g^2(t_0)X'(t_0)X^{-1}(t_0) + \int_{t_0}^{t} [(g'(s))^2 E - g^2(s)P(s)]\, ds.$$

Let λ_0 be the maximum of the absolute values of the eigenvalues of the matrices $X_i'(t_0)X_i^{-1}(t_0)$ for $i=1,2$. Then we have $g^2(t_0)X'(t_0)X^{-1}(t_0) \leqq \lambda_0 g^2(t_0)E$. Now when we let

$$A = \lambda_0 g^2(t_0)E - \int_a^{t_0} [(g'(s))^2 E - g^2(s)P(s)]\, ds,$$

(5) yields

$$(6)\qquad X'(t)X^{-1}(t) \leqq \frac{1}{g^2(t)}\left\{A + \int_a^t [(g'(s))^2 E - g^2(s)P(s)]\, ds\right\} \equiv Q(t; g, A)$$

for all $t \geqq t_0$. In view of Lemma 4, A is a linear combination of the eigenprojections of P. Since P is Hermitian, it is clear that A and thus Q are Hermitian. Moreover, X is Hermitian and conjoined and so $Q - X'X^{-1}$ is also Hermitian. By Lemma 12, $P(s)X(t) = X(t)P(s)$ and so A and thus Q also commutes with $X(t)$. Since X and X' commute, it is clear that $Q - X'X^{-1}$ will commute with X^2. Therefore for $t \geqq t_0$, (6) yields $X'X \leqq QX^2$.

Recalling that X represents either X_1 or X_2 and defining $U = X_1^2 + X_2^2$, we have that $U' = 2(X_1'X_1 + X_2'X_2) \leqq 2QU$. Since Q commutes with X and thus also X^2, we have that Q and U commute. So,

$$(7)\qquad U'(t) \leqq Q(t; g, A)U(t) + U(t)Q(t; g, A).$$

Multiplying (7) on the left by $Y_{g,A}(s, t)$, on the right by $Y_{g,A}^*(s; t)$, rearranging the terms, and making use of (3) and Lemma 9(c), we obtain $(d/dt) \cdot [Y_{g,A}(s; t)U(t)Y_{g,A}^*(s; t)] \leqq 0$. Therefore for any fixed $T \geqq t_0$ and for all $t \geqq T \geqq t_0$ and for every $s \in [a, \infty)$, integrating this last result from T to t yields

$$(8)\qquad Y_{g,A}(s; t)U(t)Y_{g,A}^*(s; t) \leqq Y_{g,A}(s; T)U(T)Y_{g,A}^*(s; T).$$

Multiplying (8) on the left by $Y_{g,A}(t; s)$, on the right by $Y_{g,A}^*(t; s)$, and using Lemma 9(a) and (b), we obtain $U(t) \leqq Y_{g,A}(t; T)U(T)Y_{g,A}^*(t; T)$ and thus $\operatorname{tr}[U^{-1}(t)] \geqq 1/\beta \operatorname{tr}\{[Y_{g,A}(t; T)Y_{g,A}^*(t; T)]^{-1}\}$, where $\beta > 0$ is the maximum eigenvalue of $U(T) > 0$. Hence, for any $b \in [T, \infty)$,

$$(9)\qquad \int_b^\infty \operatorname{tr}[U^{-1}(t)]\, dt \geqq \frac{1}{\beta}\int_b^\infty \operatorname{tr}\{[Y_{g,A}(t; T)Y_{g,A}^*(t; T)]^{-1}\}\, dt.$$

With τ as described in the statement of the theorem, Lemma 9(a) yields

$$\begin{aligned} Y_{g,A}(t; \tau)Y_{g,A}^*(t; \tau) &= Y_{g,A}(t; T)Y_{g,A}(T; \tau)Y_{g,A}^*(T; \tau)Y_{g,A}^*(t; T) \\ &\geqq \alpha Y_{g,A}(t; T)Y_{g,A}^*(t; T) \end{aligned}$$

where $\alpha > 0$ is the minimum eigenvalue of $Y_{g,A}(T; \tau)Y_{g,A}^*(T; \tau) > 0$. So, (9) leads to

$$\int_b^\infty \operatorname{tr}[U^{-1}(t)]\, dt \geqq \frac{\alpha}{\beta}\int_b^\infty \operatorname{tr}\{[Y_{g,A}(t; \tau)Y_{g,A}^*(t; \tau)]^{-1}\}\, dt,$$

and thus with c as defined in the statement of the theorem, we have

$$\begin{aligned} \int_b^\infty \operatorname{tr}[U^{-1}(t)]\, dt \geqq \frac{\alpha}{\beta}\Bigg[&\int_b^c \operatorname{tr}\{[Y_{g,A}(t; \tau)Y_{g,A}^*(t, \tau)]^{-1}\}\, dt \\ &+ \int_c^\infty \operatorname{tr}\{[Y_{g,A}(t; \tau)Y_{g,A}^*(t; \tau)]^{-1}\}\, dt\Bigg]. \end{aligned}$$

Because of the hypothesis (4), we have that $\int_b^\infty \operatorname{tr}[U^{-1}(t)]\,dt = \infty$. Since $U(t)$ as defined is simply $V^2(t)$, where $V(t)$ is a positive definite Hermitian solution of (2), we may conclude from Lemma 1 that (1) is in fact oscillatory on $[a, \infty)$.

To prove the converse, suppose the contrary; then for any choice of a C^1 function $g(t) > 0$ and $c, \tau \in [a, \infty)$ there exist scalar constants $k_1, \cdots, k_r$ with $A = \sum_{i=1}^r k_i G_i$ such that

$$\int_c^\infty \operatorname{tr}\{[Y_{g,A}(t;\tau)\,Y_{g,A}^*(t,\tau)]^{-1}\}\,dt < \infty. \tag{10}$$

Because of Lemma 4, A is a Hermitian constant matrix and also $Q(t; g, A)$ may be written as $Q(t; g, A) = \sum_{i=1}^r q_i(t; g, A) G_i$ where

$$q_i(t; g, A) = \frac{1}{g^2(t)}\left\{k_i + \int_a^t [(g'(s))^2 - \lambda_i(s) g^2(s)]\,ds\right\}.$$

Hence, the solution to (3) is expressible as $Y_{g,A}(t; T) = \sum_{i=1}^r G_i Y_{i,g,A}(t; T)$ where $Y_{i,g,A}(t; T)$ is the solution to the initial value problem

$$Y_i' = q_i(t; g, A)\,Y_i, \qquad Y_i(T) = E. \tag{11}$$

But for each $i = 1, \cdots, r$ the solution to (11) is easily seen to be

$$Y_{i,g,A}(t; T) = \exp\left(\int_T^t q_i(u; g, A)\,du\right)E,$$

and so

$$Y_{g,A}(t; T) = \sum_{i=1}^r G_i \exp\left(\int_T^t q_i(u; g, A)\,du\right). \tag{12}$$

Because of the properties of G_i specified in Lemma 4, using (12) we get

$$[Y_{g,A}(t; T)\,Y_{g,A}^*(t; T)]^{-1} = \sum_{i=1}^r G_i \exp\left(-2\int_T^t q_i(u; g, A)\,du\right),$$

and thus

$$\operatorname{tr}[Y_{g,A}(t; T)\,Y_{g,A}^*(t; T)]^{-1} = \sum_{i=1}^r \gamma_i \exp\left(-2\int_T^t q_i(u; g, A)\,du\right),$$

where $\gamma_i = \operatorname{tr} G_i$. Now $\gamma_i > 0$ since G_i is idempotent, and hence from (10) we conclude that

$$\int_c^\infty \exp\left(-2\int_T^t q_i(u; g, A)\,du\right)dt < \infty$$

for $i = 1, \cdots, r$. Now using Ráb's Hauptsatz we have that $x'' + \lambda_i(t)x = 0$ is nonoscillatory on $[a, \infty)$ for all $i = 1, \cdots, r$, and so from Corollary 8 it follows that (1) is nonoscillatory on $[a, \infty)$. □

It is appropriate at this juncture to look at a somewhat similar result in [16]. With much less restrictive assumptions on P, we can obtain from the main theorem in [16] the following.

If β is some number in the interval $(0, 2]$ and if $X'' + \beta^{-1}PX = 0$ is nonoscillatory on $[\alpha, \infty)$, then $\int_\alpha^\infty \operatorname{tr}\{[Y_B(t;\tau)\,Y_B^*(t;\tau)]^{-1}\}\,dt = \infty$, where $Y_B(t;\tau)$ satisfies $Y' = [B - \int_\alpha^t P(s)\,ds]Y$, $Y(\tau) = E$ and B is the specific constant matrix $\lim_{t\to\infty} (t-\alpha)^{-1}\int_\alpha^t \int_\alpha^s P(u)\,du\,ds$.

If in fact P in this result is assumed to have the same properties as in Theorem 13, we note that B is Hermitian and can be written as a linear combination of the eigenprojections of P. Now the contrapositive of the sufficiency portion of Theorem 13 yields the following corollary.

COROLLARY 14. *Let $P(t)$ be analytic, Hermitian, and functionally commutative on $[a, \infty)$. If (1) is nonoscillatory on $[a, \infty)$, then there exists a Hermitian constant matrix A, which is a linear combination of the eigenprojections of P such that*

$$\int_a^\infty \operatorname{tr}\{[Y_A(t;\tau)Y_A^*(t;\tau)]^{-1}\}\,dt < \infty$$

where $Y_A(t;\tau)$ satisfies $Y' = [A - \int_a^t P(s)\,ds]Y$, $Y(\tau) = E$.

A comparison of this corollary with the result stated above from [16] when $\beta = 1$, indicates that the corollary is not true for all such A.

For our final result, let us consider the maximum eigenvalue of $P(t)$. For each t, let $\lambda_{\max}(t) = \max_i \lambda_i(t)$ where we recall that $\{\lambda_i(t) \mid i = 1, \cdots, r\}$ is the set of distinct analytic eigenvalues for the analytic, Hermitian, functionally commutative matrix $P(t)$. We note that, in general, $\lambda_{\max}(t)$ will not be one of these analytic eigenvalues.

THEOREM 15. *Let $P(t)$ be analytic, Hermitian, and functionally commutative on $[a, \infty)$. If (1) is oscillatory on $[a, \infty)$, then the scalar equation $x'' + \lambda_{\max}(t)x = 0$ is* oscillatory on $[a, \infty)$.

Proof. By Corollary 8, $x'' + \lambda(t)x = 0$ is oscillatory on $[a, \infty)$ for at least one of the analytic eigenvalues $\lambda(t)$ of $P(t)$. Since $\lambda(t) \leqq \lambda_{\max}(t)$, the result follows from the Sturm comparison theorem. □

A similar result is attained in [15, p. 398].

REFERENCES

[1] J. H. BARRETT, *A Prüfer transformation for matrix differential equations*, Proc. Amer. Math. Soc., 8 (1957), pp. 510-517.

[2] G. J. BUTLER AND L. H. ERBE, *Oscillation theory for second order differential systems with functionally commutative matrix coefficients*, Funkcial. Ekvac., 28 (1985), pp. 47-55.

[3] G. J. ETGEN, *Oscillatory properties of certain nonlinear matrix differential systems of second order*, Trans. Amer. Math. Soc., 122 (1966), pp. 289-310.

[4] H. I. FREEDMAN, *Functionally commutative matrices and matrices with constant eigenvectors*, Linear and Multilinear Algebra, 4 (1976), pp. 107-113.

[5] S. GOFF, *Hermitian function matrices which commute with their derivative*, Linear Algebra Appl., 36 (1981), pp. 33-40.

[6] ———, *The Bohl transformation and oscillation of linear differential systems*, Ph.D. thesis, University of Massachusetts, Amherst, MA, 1978.

[7] S. GOFF AND D. F. ST. MARY, *The Bohl transformation for second order linear differential systems*, J. Math. Anal. Appl., to appear.

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

[9] M. MORSE, *The calculus of variations in the large*, American Mathematical Society Colloquium Publication, Vol. 18, American Mathematical Society, New York, 1934.

[10] S. PERLIS, *Theory of Matrices*, Addison-Wesley, Reading, MA, 1958.

[11] M. RÁB, *Kriterien für die Oszillation der Lösungen der Differentialgleichung* $[p(x)y']' + q(x)y = 0$, Časopis Pěst. Mat., 84 (1959), pp. 335-370; Erratum, Časopis Pěst. Mat., 85 (1960), p. 91.

[12] W. T. REID, *A Prüfer transformation for differential systems*, Pacific J. Math., 8 (1958), pp. 575-584.

[13] ———, *Ordinary differential equations*, John Wiley, New York, 1971.

[14] F. RELLICH, *Perturbation theory of eigenvalue problems*, Institute of Mathematical Sciences, New York University, New York, 1950.

[15] D. F. ST. MARY, *On transformation and oscillation of linear differential systems*, Canad. J. Math., 29 (1977), pp. 392-399.

[16] ———, *Riccati integral equations and non-oscillation of self-adjoint linear systems*, J. Math. Anal. Appl., 121 (1987), pp. 109-118.

SIAM J. MATH. ANAL.
Vol. 20, No. 1, January 1989

018

BOUNDARY LOCAL TIME AND SMALL PARAMETER EXIT PROBLEMS WITH CHARACTERISTIC BOUNDARIES*

MARTIN V. DAY†

Abstract. The exit problem for an asymptotically small random perturbation of a stable dynamical system $x(t)$ in a region D is considered. The connection between the distribution of the position of first exit and the equilibrium density of the perturbed system subject to reflection from the boundary of D is developed. Earlier work treated the case in which $x(t)$ enters D nontangentially. Here the case in which $x(t)$ is everywhere tangent to the boundary is examined. The "small-noise" asymptotics of the boundary local time turn out to be of primary importance.

Key words. exit problem, small noise, reflecting diffusion, local time

AMS(MOS) subject classifications. Primary 60H10; secondary 60J55, 60J60

1. Introduction. In this paper we will extend the results of [3], which are central to our treatment [2] of the small parameter exit problem for diffusions, to cases in which the classical assumption (1.5) fails. Our specific assumptions are given in §2. This introduction provides an overview.

Let $D \subseteq \mathbb{R}^d$ be a bounded open domain. The "exit problem" is to determine the (weak) limit (or limit points)

$$\lim_{\epsilon \downarrow 0} \mu^{\epsilon}_{x_0}(dy) = ?,$$

where $\mu^{\epsilon}_{x_0}$ is the exit distribution

$$\mu^{\epsilon}_{x_0}(dy) = P_{x_0}[x^{\epsilon}(\tau_{\partial D}) \in dy] \tag{1.1}$$

for a diffusion $x^{\epsilon}(t)$ whose generator in D is given by a (nondegenerate) second-order operator

$$\mathscr{L}^{\epsilon}u(x) = \frac{\epsilon}{2}\sum_{i,j=1}^{d} a_{ij}(x)u_{x_ix_j} + \sum_{1}^{d} b_i(x)u_{x_i}, \tag{1.2}$$

and where

$$\tau_{\partial D} = \inf\{t > 0 : x^{\epsilon}(t)\epsilon\partial D\} \tag{1.3}$$

is the first exit time from D. The cases of particular interest to us are those in which the limiting deterministic flow

$$\dot{x}^0(t) = b(x^0(t)),\ x^0(0) = x_0 \tag{1.4}$$

is in some way stable in D, with points x_0 near ∂D being repelled deeper into the interior of D as t increases.

In the most familiar version of this problem (1.4) is assumed to cross ∂D nontangentially into D:

$$\langle b(y), n(y)\rangle < 0 \quad \forall y \in \partial D. \tag{1.5}$$

*Received by the editors October 16, 1988; accepted for publication (in revised form) June 27, 1988. This research was supported in part by National Science Foundation grant DMS-8420755. The work began while the author was visiting the Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0123.

(Here and throughout $n(y)$ is the unit outward normal vector at $y \in \partial D$.) This is the famous problem discussed by Wentzell and Freidlin [6]. Culminating in [2], our approach to this problem reduced it to the asymptotic evaluation of a Laplace integral over ∂D. That reduction was based on an asymptotic connection [3] between the exit measures and the equilibrium density $p^\epsilon(x)$ for $x^\epsilon(t)$. The approach was inspired by the asymptotic calculations of Matkowsky and Schuss [12], although as noted in [2], the connection with the equilibrium density was observed in earlier work of Y. I. Kifer.

In this paper we will develop the connection between the equilibrium density and the exit problem in cases for which

$$\langle b(y), n(y)\rangle = 0 \quad \forall y \in \partial D. \tag{1.6}$$

(See (2.9) for an additional hypothesis.) In [11] and [13] the Matkowsky and Schuss approach is developed for two-dimensional examples of this type. Our probabilistic development is again strongly motivated by their work.

The treatment of this paper is incomplete in that we do not include a study of the asymptotics of p^ϵ itself. We leave that for the future. As we will see, this means we must leave our conclusions for the exit problem in a somewhat awkward form; see (6.2). We will, however, venture a conjecture which, if true, renders a considerable simplification (6.4) of our conclusions.

We pointed out in [2] that the relationship between the exit measure and equilibrium density involved the expected "local time" of $x^\epsilon(t)$ on ∂D. Because it is in fact at the heart of that relationship, we have chosen to include it explicitly in our characterization of $x^\epsilon(t)$ by taking $x^\epsilon(t)$ to be the diffusion with generator $\mathscr{L}^\epsilon$ in D and instantaneous conormal reflection off ∂D. The boundary local time occurs explicitly as the continuous, nondecreasing process $l^\epsilon(t)$ in the stochastic differential equation characterization of the reflecting diffusion $x^\epsilon(t)$, (2.13) below. Of course, given $x_0 \in D$ the presence of the reflection does not effect the exit measures $\mu^\epsilon_{x_0}$. It does, however, make $x^\epsilon(t)$ a process with compact state space $\overline{D}$, so that its equilibrium density $p^\epsilon(x)$ will exist without the modifications on D^c that were needed in [2] and [3].

The reason for the choice of conormal as the reflection direction is perhaps not apparent in the analysis below. Indeed the results below can be generalized to any nontangential reflection direction. However, there are reasons for choosing the conormal which will become apparent in subsequent work. One is that the boundary conditions satisfied by $p^\epsilon(x)$ on ∂D are simpler in the conormal case, so that future study of $p^\epsilon(x)$ will be more convenient in that context. Secondly, large deviations results for the reflected process $x^\epsilon(t)$ are simpler in the conormal case.

Section 2 presents our technical assumptions and notation. Sections 3-6 contain the main arguments relating to the exit problem, with the more distracting technical arguments and proofs deferred to §7-9. In §3 we will derive the "fundamental equation" (3.7) that connects the exit measures μ^ϵ with the equilibrium density p^ϵ via an operator B^ϵ on $C(\partial D)$. In §4 we will see that the asymptotic behavior of B^ϵ is given by $B^\epsilon \sim \epsilon^{-1/2}B$, where B is another operator on $C(\partial D)$. To deduce the limit points of μ^ϵ we need to invert B in the limiting form of the fundamental equation. Explicit formulas for B and B^{-1} are obtained in §5.

In §6 we will exhibit the resulting formula for the limit points of μ^ϵ and conclusions for the exit problem. Also in §6 is the conjecture regarding asymptotic behavior of p^ϵ mentioned above. Section 7 contains some estimates that will be used in the proof of Theorem 2.1 given in §8 and the proof of Theorem 4.1 given in §9.

2. Assumptions and preliminaries. Here we will outline our technical assumptions. First, regarding regularity of the coefficient functions and ∂D, we assume the following. The $\mathbb{R}^d$ valued function $b(\cdot)$ is in $C^{1,\lambda}(\overline{D})$. The $d\times d$ matrix valued function $\sigma(x)$ is $C^{2,\lambda}(\overline{D})$ and nonsingular everywhere in $\overline{D}$; $a(x) = \sigma(x)\sigma(x)^T$. $D \subseteq \mathbb{R}^d$ is bounded, open, connected with ∂D being C^3. (The notation $C^{n,\lambda}(\overline{D})$ refers to functions which, together with all partials of order $\leq n$, are Hölder continuous with exponent $0 < \lambda \leq 1$.)

One consequence of the above is that

$$\rho(x) = \operatorname{dist}(x, \partial D) \tag{2.1}$$

is then C^3 in a narrow (closed) strip G adjoining ∂D:

$$G = \{x \epsilon \overline{D} : 0 \leq \rho(x) \leq h_0\}.$$

Several additional restrictions on h_0 will be imposed below to insure various properties of G used throughout the paper. For instance, provided h_0 is sufficiently small, for $x \in G$ we can write

$$\nabla\rho(x) = -n(y(x)) \tag{2.2}$$

where $x \to y(x)$ is a C^2 mapping of G to ∂D with $y(x) = x$ for $x \in \partial D$, and $n(y)$ is the unit outward normal to D at $y \in \partial D$. Thus

$$n(x) = -\nabla\rho(x) \tag{2.3}$$

provides a C^2 extension of the normal vector field from ∂D to G. The conormal vector field

$$\eta(x) = a(x)n(x)$$

is thereby defined on G also.

Define the positive $C(G)$ function $\alpha(\cdot)$ by

$$\alpha(x)^2 = \langle n(x), a(x)n(x)\rangle = \langle n(x), \eta(x)\rangle. \tag{2.4}$$

The following bounds will appear in our analysis:

$$A_0 = \inf_G \frac{1}{2}\sum_{i,j} a_{ij}\rho_{x_i x_j}, \tag{2.5}$$

$$A_1 = \sup_{x\in\overline{D},|z|=1} \langle z, a(x)z\rangle, \tag{2.6}$$

$$\alpha_0^2 = \inf_G \alpha(x)^2 > 0. \tag{2.7}$$

Note that

$$\sup_G \alpha(x)^2 \leq A_1. \tag{2.8}$$

The solution of (1.4) with $x^0(0) = x_0$ will be denoted $x^0(t)$, or $x^0(t; x_0)$ when we want the dependence on the initial condition to be explicit. As indicated in the introduction, we assume that

$$\langle b(y), n(y)\rangle = 0 \ \forall\ y \in \partial D.$$

This means that ∂D is an invariant set for (1.4); i.e., $x^0(t; y) \in \partial D$ whenever $y \in \partial D$. Our analysis depends heavily on the assumption that ∂D is repelling for (1.4) with

x_0 in D. Specifically, we assume that there exists a positive function $b_1 \in C^1(\partial D)$ so that

$$\lim_{x \to y} \frac{-1}{\rho(x)} \langle b(x), n(x) \rangle = b_1(y) \tag{2.9}$$

uniformly over $y \in \partial D$. Using the projection $y(x) : G \to \partial D$ as above, we can extend b_1 to a strictly positive $C^1(G)$ function by

$$b_1(x) = b_1(y(x)).$$

We can express (2.9) as

$$b(x) = -\rho(x) b_1(x) n(x) + b_0(x) + o(\rho(x)), \text{ as } \rho(x) \to 0, \tag{2.10}$$

where $b_0(x)$ is $C(G)$ with $\langle b_0(x), n(x) \rangle = 0$, for all $x \in G$. Provided h_0 is sufficiently small,

$$K = \inf_{x \in G} \frac{-\langle b(x), n(x) \rangle}{\rho(x)} \tag{2.11}$$

will be a positive constant. In particular

$$\langle n, b \rangle \leq 0 \quad \text{everywhere in } G. \tag{2.12}$$

We take $(\Omega, \mathscr{F}, P)$ to be a complete probability space with a right continuous nondecreasing family $\{\mathscr{F}_t\}_{t \geq 0}$ of complete sub σ-algebras of $\mathscr{F}$. $w(t)$ is a d-dimensional $\mathscr{F}_t$ adapted Brownian motion. For a given $x_0 \in \overline{D}$ we take $x^\epsilon(t)$ and $l^\epsilon(t)$ to be the continuous $\mathscr{F}_t$ adapted processes with $l^\epsilon(\cdot) \in [0, \infty)$ nondecreasing and $x^\epsilon(\cdot) \in \overline{D}$ which satisfy

$$dx^\epsilon(t) = b(x^\epsilon(t))dt + \epsilon^{1/2} \sigma(x^\epsilon(t))dw(t) - \frac{\epsilon}{2} \eta(x^\epsilon(t)) dl^\epsilon(t), \; x^\epsilon(0) = x_0 \tag{2.13}$$

$$dl^\epsilon(t) = \chi_{\partial D}(x^\epsilon(t)) dl^\epsilon(t), \; l^\epsilon(0) = 0.$$

The existence of a unique solution pair $(x^\epsilon(\cdot), l^\epsilon(\cdot))$ is demonstrated in [1] or [10]. As is typical, we will let P_{x_0} denote the distribution of the solution $(x^\epsilon(\cdot), l^\epsilon(\cdot))$, i.e., the induced probability measure on

$$C([0, \infty); \overline{D}) \times C([0, \infty); [0, \infty)).$$

The uniqueness of solutions to (2.13) implies the strong Markov property.

The exit time $\tau_{\partial D}$ and exit measures $\mu^\epsilon_{x_0}(dy)$ on ∂D are defined as in (1.3) and (1.1) above. In addition Γ will denote the "inne" boundary of G and, for $x_0 \epsilon G, \tau_\Gamma$ its hitting time:

$$\Gamma = \{x \epsilon D : \rho(x) = h_0\},$$
$$\tau_\Gamma = \inf\{t > 0 : x^\epsilon(t) \epsilon \Gamma\}.$$

With regard to stability of (1.4) in D, the only feature that we will actually use is the following "leveling" property, which we therefore assume:

$$\text{for any compact } \; K \subseteq D \; \text{ and } \; f \in C(\partial D)$$

$$\sup_{x,y \in K} \left| \int_{\partial D} f(z) \mu^\epsilon_x(dz) - \int_{\partial D} f(z) \mu^\epsilon_y(dz) \right| \to 0 \quad \text{as } \epsilon \downarrow 0. \tag{2.14}$$

This holds for instance if D is a domain of attraction for an isolated exponentially stable critical point of (1.4); see [4]. Equation (2.14) will also hold for some other

types of stability. Essentially it says that the set of (weak) limit points of μ_x^ϵ as $\epsilon \downarrow 0$ is independent of $x \epsilon D$.

Next we want to indicate briefly how the existence of a continuous stationary density follows from our assumptions. The results of Sato and Ueno [14] imply the existence of a transition density $p^\epsilon(x;t,y)$ which is continuous on $(0,\infty)\times\overline{D}\times\overline{D}$. The nondegeneracy of the diffusion matrix $a(\cdot)$ and compactness of $\overline{D}$ imply that $x^\epsilon(t)$ is positive recurrent, satisfying the condition (B) of Hasminskii [7, p. 118]. Theorems 4.1 and 5.1 of [7] imply that $x^\epsilon(t)$ has a unique stationary distribution, $\pi^\epsilon(dx)$. It follows then that $\pi^\epsilon(dx)$ has a continuous density $p^\epsilon(x)$ on $\overline{D}$ given by

$$p^\epsilon(x) = \int_{\overline{D}} p^\epsilon(z;1,x)\pi^\epsilon(dz).$$

The following properties of the nondecreasing process $l^\epsilon(t)$ can be deduced from the fact that $x^\epsilon(t) \in \overline{D}$ for all $t \geq 0$ and standard features of the Brownian term in (2.13):

$$l^\epsilon(t) \to +\infty \quad \text{a.s. as } t \to +\infty;$$

$$\text{if } x_0 \in \partial D, \text{ then } l^\epsilon(t) > 0 \quad \text{a.s. } \forall \ t > 0.$$

Moreover, for any $T < \infty$

$$\int_0^T \chi_{\partial D}(x^\epsilon(t))dt = 0 \quad \text{a.s.} \tag{2.15}$$

(See [1].)

$l^\epsilon(t)$ is called the "local time on ∂D" since it is a sort of spatial density for the time spent by $x^\epsilon(t)$ on ∂D, roughly analogous to the classical Brownian local time. The following theorem tailors a statement of this fact that to our needs below. Certainly a stronger statement is true. In particular we expect the convergence to be almost sure since it is in the purely Brownian case; see Hsu [8]. We defer the proof to §8.

THEOREM 2.1. *Let $x^\epsilon(0) \in G$ be given any initial distribution, independent of $w(\cdot)$. For any $f \in C(\partial D)$*

$$\lim_{h\downarrow 0} E\left[\frac{1}{h}\int_0^{\tau_\Gamma} \chi_{[0,h]}(\rho(x^\epsilon(t))f(x^\epsilon(t))dt\right] = E\left[\int_0^{\tau_\Gamma} f(x^\epsilon(t))dl^\epsilon(t)\right].$$

3. The fundamental equation for the exit measure. We will now derive (3.7), which describes the connection between the exit measures $\mu_\cdot^\epsilon(dy)$ and the stationary density $p^\epsilon(x)$. The derivation is much the same as in [3]. Assumption (2.9) is not used at all here, but (2.14) is fundamental. At the end of the section we will indicate a second way to understand the equation in terms of the "boundary process" associated with the reflecting diffusion.

The starting point is the following construction of Hasminskii [7] for the stationary distribution $\pi^\epsilon(dx) = p^\epsilon(x)dx$. Define a double sequence of stopping times by

$$\tau_{\partial D}^0 = 0$$

and recursively for $n \geq 1$,

$$\tau_\Gamma^n = \inf\{t > \tau_{\partial D}^n : x^\epsilon(t) \in \Gamma\}$$
$$\tau_{\partial D}^n = \inf\{t > \tau_\Gamma^{n-1} : x^\epsilon(t) \in \partial D\}.$$

The nondegeneracy of $a(\cdot)$ implies that all of these are almost surely finite. The $x^\epsilon(\tau^n_{\partial D})$ form a Markov chain on ∂D for which there is a unique stationary distribution $\mu^\epsilon(dy)$. The stationary distribution $p^\epsilon(x)dx$ is expressed in terms of μ^ϵ as a (normalized) expected path integral over one "cycle" of the chain from ∂D at $t = \tau^n_{\partial D}$ to Γ at $t = \tau^n_\Gamma$ and back to ∂D at $t = \tau^{n+1}_{\partial D}$, i.e., over $t \in (\tau^n_{\partial D}, \tau^{n+1}_{\partial D}]$. Specifically in terms of $n = 0$,

$$\int_D g(x)p^\epsilon(x)dx = c^\epsilon E_{\mu^\epsilon}\left[\int_0^{\tau^1_{\partial D}} g(x^\epsilon(s))ds\right] \tag{3.1}$$

for any bounded measurable g on $\overline{D}$. The normalizing constant is $c^\epsilon = E_{\mu^\epsilon}[\tau^1_{\partial D}]^{-1}$. ($E_\mu[\cdot]$ means $\int E_x[\cdot]\mu(dx)$.) For our purposes it is advantageous to break the cycle into its two halves: $[\tau^n_{\partial D}, \tau^n_\Gamma] \cup [\tau^n_\Gamma, \tau^{n+1}_{\partial D}]$. Define the measure $\lambda^\epsilon(dz)$ on Γ by

$$\lambda^\epsilon(A) = P_{\mu^\epsilon}[x^\epsilon(\tau^0_\Gamma) \in A], A \subseteq \Gamma \quad \text{measurable.}$$

It follows that for measurable $A \subseteq \partial D$

$$\mu^\epsilon(A) = E_{\lambda^\epsilon}[x^\epsilon(\tau_{\partial D}) \in A] = \int_\Gamma \mu^\epsilon_z(A)\lambda^\epsilon(dz).$$

The leveling assumption (2.14) implies that convergence as $\epsilon \downarrow 0$ of μ^ϵ is the same as for any $\mu^\epsilon_{x_0}$.

LEMMA 3.1. *The weak limit points, as $\epsilon \downarrow 0$, of μ^ϵ and $\mu^\epsilon_{x_0}$ agree and are independent of $x_0 \in D$.*

Proof. If $\{\epsilon_n\}_1^\infty$ is a sequence decreasing to 0, then for any $x_0 \in D$ and $f \in C(\partial D)$,

$$\begin{aligned}\left|\int_{\partial D} f d\mu^\epsilon - \int_{\partial D} f d\mu^\epsilon_{x_0}\right| &\le \int_\Gamma \left|\int_{\partial D} f d\mu^\epsilon_z - \int_{\partial D} f d\mu^\epsilon_{x_0}\right| \lambda^\epsilon(dz) \\ &\le \sup_{z \in \Gamma \cup \{x_0\}} \left|\int f d\mu^\epsilon_z - \int f d\mu^\epsilon_{x_0}\right|\end{aligned}$$

which converges to 0 by (2.14). □

Our goal then is to identify the limit points of μ^ϵ in terms of p^ϵ as $\epsilon \downarrow 0$. Using the two halves of the cycle, we can rewrite (3.1) as

$$\int_D g(x)p^\epsilon(x)dx = c^\epsilon\left(E_{\mu^\epsilon}\left[\int_0^{\tau_\Gamma} g(x^\epsilon(s))ds\right] + E_{\lambda^\epsilon}\left[\int_0^{\tau_{\partial D}} g(x^\epsilon(s))ds\right]\right). \tag{3.2}$$

Now we perform a sort of spatial differentiation in (3.2). To be precise, let $f \in C(\partial D)$. Extend f to $f \in C(G)$, and for $0 < h < h_0$ define

$$f^{(h)}(x) = h^{-1}\chi_{[0,h]}(\rho(x))f(x).$$

Use this in (3.2) and pass to the limit as $h \downarrow 0$. By virtue of the continuity of p^ϵ we get a surface integral on the left:

$$\lim_{h \downarrow 0} \int_D \frac{1}{h}\chi_{[0,h]}(\rho(x))f(x)p^\epsilon(x)dx = \int_{\partial D} f(y)p^\epsilon(y)dy. \tag{3.3}$$

According to Corollary 1 of [3], the second term on the right vanishes:

$$\lim_{h \downarrow 0} E_{\lambda^\epsilon}\left[\int_0^{\tau_{\partial D}} f^{(h)}(x^\epsilon(s))ds\right] = 0 \tag{3.4}$$

(The notation here is slightly different from [3]; the λ^ϵ and $\tau_{\partial D}$ used here were denoted by ν^ϵ and τ_D there.) The convergence of the first term on the right in (3.2) is given by Theorem 2.1 above. We find then that for all $f \epsilon C(\partial D)$ the following equation holds:

$$\int_{\partial D} f(y)p^\epsilon(y)dy = c^\epsilon \int_{\partial D} E_y \left[\int_0^{\tau_\Gamma} f(x^\epsilon(t))dl^\epsilon(t)\right] \mu^\epsilon(dy). \tag{3.5}$$

Define the "local time operator" B^ϵ on $C(\partial D)$ by

$$B^\epsilon[f](y) = E_y \left[\int_0^{\tau_\Gamma} f(x^\epsilon(t))dl^\epsilon(t)\right].$$

Also define the measure ν^ϵ on ∂D by

$$\nu^\epsilon(dy) = c^\epsilon p^\epsilon(y)dy, \tag{3.6}$$

where c^ϵ is a new normalizing constant to make ν^ϵ a probability measure. (We will use c^ϵ as a generic ϵ-dependent normalizing constant; its actual value may vary from one equation to the next.) Equation (3.5) may now be written

$$\int_{\partial D} f(y)\nu^\epsilon(dy) = c^\epsilon \int_{\partial D} B^\epsilon[f](y)\mu^\epsilon(dy) \quad \forall f \epsilon C(\partial D). \tag{3.7}$$

We will call (3.7) the "fundamental equation" for the exit measure because it describes precisely the connection between μ^ϵ and p^ϵ. It is true that both μ^ϵ and B^ϵ are dependent on the choice of G through the hitting times τ_Γ^n. However, in the limit as $\epsilon \downarrow 0$ this dependence vanishes. (For μ^ϵ this is a consequence of the leveling property; for B^ϵ it follows from the asymptotic formula of the next section.) The goal of the subsequent sections is to invert B^ϵ in the limit as $\epsilon \downarrow 0$ so that the limit points of μ^ϵ can be determined from those of ν^ϵ.

There is a second way to understand our fundamental equation, based on what is called the boundary process $y^\epsilon(l)$ associated with $x^\epsilon(t)$. Since none of our proofs will be based on this interpretation, we will limit ourselves to a summary discussion. (See [14] and [15] for more on the boundary process.)

The boundary process can be obtained from $x^\epsilon(t)$ by making a random time change so that $l = l^\epsilon(t)$ becomes the independent time parameter, rather than the "real" time t. This is done by defining the right continuous inverse $t_+^\epsilon(l)$ of $l^\epsilon(t)$:

$$t_+^\epsilon(l) = \sup\{t > 0 : l^\epsilon(t) \leq l\}.$$

The boundary process is given by

$$y^\epsilon(l) = x^\epsilon(t_+^\epsilon(l)).$$

This turns out to be a right continuous strong Markov process on ∂D, moving only by discontinuous jumps. Also,

$$\eta^n = l^\epsilon(\tau_{\partial D}^n)$$

form a sequence of stopping times for $y^\epsilon(l)$ (using the σ-algebras $\mathscr{F}_{t_+^\epsilon(l)}$) and

$$y^\epsilon(\eta^n) = x^\epsilon(\tau_{\partial D}^n).$$

As before, μ^ϵ is the stationary distribution for this Markov chain. We can now construct a stationary distribution $\bar{\nu}^\epsilon$ for $y^\epsilon(l)$ analogously to (3.1):

$$\int_{\partial D} f(y)\bar{\nu}^\epsilon(dy) = c^\epsilon E_{\mu^\epsilon}\left[\int_0^{\eta^1} f(y^\epsilon(l))dl\right]. \tag{3.8}$$

However we may check that

$$\begin{aligned} E_y\left[\int_0^{\eta^1} f(y^\epsilon(l))dl\right] &= E_y\left[\int_0^{\tau^1_{\partial D}} f(x^\epsilon(t))dl^\epsilon(t)\right] \\ &= E_y\left[\int_0^{\tau_\Gamma} f(x^\epsilon(t))dl^\epsilon(t)\right] \\ &= B^\epsilon[f](y). \end{aligned}$$

Thus (3.8) becomes

$$\int_{\partial D} f(y)\bar{\nu}^\epsilon(dy) = c^\epsilon \int_{\partial D} B^\epsilon[f](y)\mu^\epsilon(dy), \quad \forall f \in C(\partial D).$$

We recognize this as our fundamental equation (3.7). What we learn is that ν^ϵ defined by (3.6) is in fact a stationary distribution $\bar{\nu}^\epsilon$ for $y^\epsilon(l)$, and that (3.7) can be thought of as its construction in terms of the embedded Markov chain. We should point out that this connection between the stationary density of a reflected diffusion and the stationary distribution of its boundary process has been noted before by Freidlin [6, p. 174].

4. Asymptotic evaluation of B^ϵ. We now consider the behavior of the boundary local time operator B^ϵ as $\epsilon \downarrow 0$. The asymptotic behavior is revealed by considering the one-dimensional process $\varsigma^\epsilon(t)$ defined for $0 \le t \le \tau_\Gamma$ by

$$\varsigma^\epsilon(t) = \epsilon^{-1/2}\rho(x^\epsilon(t)). \tag{4.1}$$

The asymptotic properties of $\varsigma^\epsilon(t)$ give a "boundary layer" description of $x^\epsilon(t)$ which is roughly analogous to the boundary layer calculations in [12], [13]. Note that

$$\begin{aligned} \tau_\Gamma &= \inf\{t > 0 : x^\epsilon(t) \in \Gamma\} \\ &= \inf\{t > 0 : \rho(x^\epsilon(t)) = h_0\} \\ &= \inf\{t > 0 : \varsigma^\epsilon(t) = \epsilon^{-1/2}h_0\} \\ &= \eta^\epsilon_{\epsilon^{-1/2}h_0}, \end{aligned} \tag{4.2}$$

where in general for $\epsilon^{1/2}\varsigma^\epsilon(0) = \rho(x^\epsilon(0)) < r < h_0$ we will write

$$\eta^\epsilon_r = \inf\{t > 0 : \varsigma^\epsilon(t) = r\}. \tag{4.3}$$

Using the fact that $\langle \nabla\rho(y), \eta(y)\rangle = -\alpha(y)^2$ on ∂D, Itô's formula (see [9, p. 66]) gives

$$d\varsigma^\epsilon(t) = \epsilon^{-1/2}\mathscr{L}^\epsilon\rho(x^\epsilon(t))dt + \langle \sigma^T\nabla\rho(x^\epsilon(t)), dw(t)\rangle + \frac{1}{2}\epsilon^{1/2}\alpha(x^\epsilon(t))^2 dl^\epsilon(t), \tag{4.4}$$

with $l^\epsilon(t)$ increasing only when $\varsigma^\epsilon(t) = 0$.

Now if $\varsigma \in [0,\infty)$ and $x \in G$ are variables related by $\varsigma = \epsilon^{-1/2}\rho(x)$, then (2.10) implies that

$$\begin{aligned} \epsilon^{-1/2}\mathscr{L}^\epsilon\rho(x) &= \frac{\epsilon^{1/2}}{2}\sum_{i,j} a_{ij}(x)\rho_{x_ix_j}(x) + \epsilon^{-1/2}\langle \nabla\rho(x), b(x)\rangle \\ &= \epsilon^{-1/2}\langle \nabla\rho(x), b(x)\rangle + O(\epsilon^{1/2}) \\ &= \epsilon^{-1/2}\langle -n(x), -\rho(x)b_1(x)n(x)\rangle + O(\epsilon^{1/2})\rangle + \epsilon^{-1/2}o(\rho) \\ &= b_1(x)\varsigma + o(1), \end{aligned} \tag{4.5}$$

the $o(1)$ being uniform for ς in compact sets, as $\epsilon \downarrow 0$.

Using $\alpha(\cdot)$ as in (2.4), the second term of (4.4) is of the form

$$\langle \sigma^T(x^\epsilon(t))n(x^\epsilon(t)), dw(t)\rangle = \alpha(x^\epsilon(t))d\beta(t), \tag{4.6}$$

where $\beta(t)$ is a one-dimensional Brownian motion adapted to the $\mathscr{F}_t$. To be precise, given $\epsilon > 0$, (4.6) defines an $\mathscr{F}_t$ adapted process $\beta(t)$. (To extend the definition to $t > \tau_\Gamma$, replace $x^\epsilon(t)$ by $x^\epsilon(t \wedge \tau_\Gamma)$ on both sides.) It is simple to check that $\beta(t)$ is a Brownian motion using $\|\sigma^T n\| = \alpha$ and a martingale characterization, such as [9, Thm. II-6.1]. (Although this definition of β is ϵ-dependent, we ignore this dependence in our present heuristic discussion. For our purposes this can be justified because we are only concerned with the distributions of $\varsigma^\epsilon(\cdot)$ and $\varsigma(\cdot)$ below. See the paragraph preceding Lemma 9.2 for more on this point.)

As $\epsilon \downarrow 0$, $x^\epsilon(\cdot) \to x^0(\cdot)$. (We will prove this in §9 below). Thus if $x^\epsilon(0) = y \in \partial D$, we would expect that $\varsigma^\epsilon(t)$ converges in distribution to the one-dimensional reflecting diffusion $\varsigma(t)$ on $[0, \infty)$ described by

$$d\varsigma(t) = b_1(x^0(t))\varsigma(t)dt + \alpha(x^0(t))d\beta(t) + \frac{1}{2}\alpha(x^0(t))^2 dl(t),\ \varsigma(0) = \varsigma_0 \tag{4.7}$$

$$dl(t) = \chi_{\{0\}}(\varsigma(t))dl(t),\ l(0) = 0 \tag{4.8}$$

$$\dot{x}^0(t) = b(x^0(t)),\ x^0(0) = y, \tag{4.9}$$

with $\varsigma_0 = \varsigma^\epsilon(0) = 0$. As usual, this is to be solved simultaneously for the pair of continuous processes $(\varsigma(t), l(t))$ under the restrictions that $l(t)$ is nondecreasing and $\varsigma(t) \geq 0$. Probabilities and expectations with respect to (4.7), (4.8), and (4.9) will be indicated using $P_{\varsigma_0, y}$ and $E_{\varsigma_0, y}$.

Comparing the local time terms in (4.4) and (4.7), we would also expect that $\epsilon^{1/2} l^\epsilon(t)$ converges to $l(t)$. Moreover, (4.2) suggests that $\tau_\Gamma \to +\infty$ almost surely as $\epsilon \downarrow 0$. Thus we anticipate that

$$\begin{aligned} B^\epsilon[f](y) &= E_y\left[\int_0^{\tau_\Gamma} f(x^\epsilon(t))dl^\epsilon(t)\right] \\ &= \epsilon^{-1/2} E_y\left[\int_0^{\tau_\Gamma} f(x^\epsilon(t))d(\epsilon^{1/2} l^\epsilon(t))\right] \\ &\sim \epsilon^{-1/2} E_{0,y}\left[\int_0^{\infty} f(x^0(t))dl(t)\right]. \end{aligned}$$

Define the operator B on $C(\partial D)$ by

$$B[f](y) = E_{0,y}\left[\int_0^{\infty} f(x^0(t))dl(t)\right].$$

Note that $y \in \partial D$ determines the solution $x^0(t)$ of (4.9), which is a deterministic trajectory on ∂D. This path then determines the time dependent coefficients in the equation for $\varsigma(t)$, from which we get $l(t)$, which in turn provides a weighting of points on $x^0(\cdot)$. Thus $B[f](y)$ is a certain integral of the values of f along the solution of (4.9) with $x^0(0) = y$.

The above considerations suggest the following result.

THEOREM 4.1. *Given any $f \in C(\partial D)$, $\epsilon^{1/2} B^\epsilon[f]$ converges to $B[f](y)$ uniformly on ∂D.*

A fair amount of technical work is needed to make the above discussion into a proof of this theorem. We therefore postpone the proof until §9, so as not to interrupt the continuity of our consideration of the exit problem itself.

5. A canonical time scale and calculation of $B[f]$. We will now see that a change of variables uncouples the equations (4.7) and (4.9) and allows the derivation of explicit formulas for $B[f]$ and $B^{-1}[f]$. The idea is to select a positive function $\gamma(\cdot) \in C(\partial D)$ and change from ς, t to new variables ξ, s according to

$$\xi(s) = \gamma(x^0(t))\varsigma(t), \quad s = \int_0^t \alpha(x^0(u))^2\gamma(x^0(u))^2 du.$$

In these variables (4.7) becomes

$$\begin{aligned} d\xi(s) = \alpha(x^0(t))^{-2}\gamma(x^0(t))^{-2}\left[b_1(x^0(t)) + \gamma(x^0(t))^{-1}\frac{d}{dt}\gamma(x^0(t))\right]\xi(s)ds \\ +d\bar{\beta}(s) + \frac{1}{2}\gamma(x^0(t))\alpha(x^0(t))^2 dl(t), \end{aligned} \tag{5.1}$$

where $\bar{\beta}(s) = \int_0^t \alpha\gamma(x^0(u))d\beta(u)$ is a Brownian motion on the s time scale. We want to choose $\gamma(\cdot) > 0$ so that

$$\alpha^{-2}\gamma^{-2}\left[b_1(x^0) + \gamma^{-1}\frac{d}{dt}\gamma(x^0)\right] = 1 \text{ everywhere on } \partial D, \tag{5.2}$$

or

$$\frac{d}{dt}\gamma(x^0(t)) + b_1(x^0)\gamma(x^0) = \alpha^2(x^0)\gamma(x^0)^3.$$

LEMMA 5.1. *There exists a unique positive function $\gamma \in C(\partial D)$, differentiable along every trajectory of $\dot{x}^0 = b(x^0)$ on ∂D, satisfying* (5.2).

By "differentiable along every trajectory" we mean that for every $y = x^0(0) \in \partial D$, $\gamma(x^0(t))$ is differentiable in t. This is weaker than saying simply that $\gamma(\cdot)$ is differentiable, which may in fact be false.

Proof. Note that the change of variable $Z(y) = \gamma(y)^{-2}$ makes (5.2) linear:

$$-\frac{1}{2}\frac{d}{dt}Z(x^0) + b_1(x^0)Z(x^0) = \alpha(x^0)^2.$$

A simple integrating factor calculation shows that, for any such $Z(\cdot)$ and any solution $x^0(t)$ of $\dot{x}^0 = b(x^0)$ on ∂D,

$$Z(x^0(0)) = Z(x^0(T))\exp\left[-2\int_0^T b_1(x^0(t))dt\right] + 2\int_0^T \alpha(x^0(t))^2 \exp\left[-2\int_0^t b_1(x^0(u))du\right]dt$$

for all $T > 0$. Since $b_1 > 0$ on ∂D, the first term vanishes as $T \to \infty$, showing that Z is given uniquely by

$$Z(y) = 2\int_0^\infty \alpha(x^0(t))^2 \exp\left[-2\int_0^t b_1(x^0(u))du\right]dt,$$

where $x^0(0) = y \in \partial D$. This is easily argued to be positive, continuous on ∂D, and differentiable along any trajectory $x^0(t)$. □

This choice of $\gamma(\cdot)$ simplifies (5.1) to

$$d\xi(s) = \xi(s)ds + d\bar{\beta}(s) + \frac{1}{2}\gamma(x^0(t))\alpha(x^0(t))^2 dl(t).$$

It is natural to complete this new system of variables by defining a rescaled local time,

$$\bar{l}(s) = \int_0^t \gamma(x^0(u))\alpha(x^0(u))^2 dl(u).$$

Equation (4.9) can be recast in terms of s as

$$\frac{d}{ds}x^0(s) = \bar{b}(x^0(s)), \tag{5.3}$$

where $\bar{b}$ is defined in ∂D by

$$\bar{b}(y) = \frac{b(y)}{\alpha(y)^2\gamma(y)^2}.$$

Thus in the new variables (4.7), (4.8), and (4.9) become

$$\begin{aligned} &d\xi(s) = \xi(s)ds + d\bar{\beta}(s) + \frac{1}{2}d\bar{l}(s), \ \xi(0) = 0 \\ &\bar{l}(s) = \int_0^s \chi_{\{0\}}(\xi(v))d\bar{l}(v) \\ &\frac{d}{ds}x^0(s) = \bar{b}(x^0(s)), \ x^0(0) = y \end{aligned} \tag{5.4}$$

with the usual nonnegativity and nondecreasing requirements for $\xi(\cdot)$ and $\bar{l}(\cdot)$, respectively.

This change of variables has accomplished two things. One is that the ξ and x^0 equations are uncoupled, and the other is that the ξ equation is linear with constant coefficients. In fact, a solution is given by $\xi(t) = |z(t)|$, where $z(t)$ is the unreflected diffusion in $\mathbb{R}^1$ with generator $\frac{1}{2}\left(\frac{d}{dz}\right)^2 + z\frac{d}{dz}$, an (unstable) Ornstein-Uhlenbeck process. This allows us to carry out some explicit calculations for $\xi(t)$. In particular the transition density for $\xi(\cdot)$ is given by

$$q(\xi_0; s, \xi_1) = [\pi\theta^2(s)]^{-1/2}[e^{-(\xi_1+e^s\xi_0)^2/\theta^2(s)} + e^{-(\xi_1-e^s\xi_0)^2/\theta^2(s)}], \tag{5.5}$$

where

$$\theta(s) = (e^{2s} - 1)^{1/2}.$$

This may be checked for instance by verifying that q is the fundamental solution of the backward PDE associated with $\xi(\cdot)$. We leave the details to the reader. We may also check that $\xi(t)$ is nonexploding, and so is defined for all $0 \le t < \infty$.

We are now in a position to derive an explicit formula for the operator B. The first step is to express it in the new variables.

$$\begin{aligned} B[f](y) &= E_{0,y}\left[\int_0^\infty f(x^0(t))dl(t)\right] \\ &= E_0\left[\int_0^\infty \frac{f}{\gamma\alpha^2}(x^0(s))d\bar{l}(s)\right], \ x^0(0) = y \\ &= \overline{B}\left[\frac{f}{\gamma\alpha^2}\right](y), \end{aligned}$$

where

$$\overline{B}[g](y) = E_0\left[\int_0^\infty g(x^0(s;y))d\bar{l}(s)\right].$$

(The notation $x^0(s;y)$ indicates the solution of (5.3) with $x^0(0) = y$). Now if $y \in \partial D$ and $g \in C(\partial D)$, then $g(x^0(s;y))$ is a bounded continuous function of $s \in [0,\infty)$. We can describe $\overline{B}$ in terms of the following operator $\mathscr{B}$ on $C_b[0,\infty)$:

$$\mathscr{B}[\varphi](s) = E_0\left[\int_0^\infty \varphi(s+v)d\bar{l}(v)\right].$$

That $\mathscr{B}[\varphi] \in C_b[0,\infty)$ follows from the explicit formula of the next lemma. $\overline{B}$ is reconstructed from $\mathscr{B}$ according to

$$\overline{B}[g](x^0(s;y)) = \mathscr{B}[\varphi](s), \text{ where } \varphi(\cdot) = g(x^0(\cdot;y)).$$

LEMMA 5.2. *For any* $\varphi \in C_b[0,\infty)$,

$$\mathscr{B}[\varphi](s) = \int_0^\infty \varphi(s+v)\frac{2}{\sqrt{\pi}}(e^{2v}-1)^{-1/2}dv.$$

Proof. The key is to calculate $E_0[\bar{l}(s)]$. By virtue of (5.4), for $\xi(0) = 0$ we have

$$\frac{1}{2}\bar{l}(s) = \xi(s) - \int_0^s \xi(v)dv - \overline{\beta}(s) \tag{5.5}$$

and so

$$\begin{aligned} E_0[\bar{l}(s)] &= 2E_0\left[\xi(s) - \int_0^s \xi(v)dv\right] \\ &= 2\{E_0[\xi(s)] - \int_0^s E_0[\xi(v)]dv\}. \end{aligned} \tag{5.6}$$

Using the density (5.5),

$$E_0[\xi(s)] = \int_0^\infty zq(0;s,z)dz = \frac{2}{\sqrt{\pi}}\int_0^\infty \frac{z}{\theta(s)}e^{-z^2/\theta^2(s)}dz = \frac{1}{\sqrt{\pi}}\theta(s).$$

Using this in (5.6), we find

$$E_0[\bar{l}(s)] = \frac{2}{\sqrt{\pi}}\left[\theta(s) - \int_0^s \theta(v)dv\right].$$

Now one easily checks that $\theta'(s) = \theta(s) + 1/\theta(s)$, with $\theta(0) = 0$, so that the preceding can be rewritten as

$$E_0\left[\int_0^s d\bar{l}(v)\right] = \frac{2}{\sqrt{\pi}}\int_0^s \theta(v)^{-1}dv.$$

This implies that

$$E_0\left[\int_0^\infty \varphi(v)d\bar{l}(v)\right] = \frac{2}{\sqrt{\pi}}\int_0^\infty \varphi(v)\theta(v)^{-1}dv \tag{5.7}$$

for φ piecewise constant with compact support. Since $\varphi \in C_b[0,\infty)$ can be uniformly approximated on compacts by piecewise constant functions, and since $\theta(v)^{-1} \in L^1[0,\infty)$, (5.7) follows for all $\varphi \in C_b[0,\infty)$. The lemma is just (5.7) applied to a translation of φ. □

The next lemma tells us how to invert $\mathscr{B}$.

LEMMA 5.3. $\mathscr{B}$ *is 1-1 on* $C_b[0,\infty)$. *If* $\psi \in C_b^1[0,\infty)$, *then* $\psi = \mathscr{B}[\varphi]$ *where* $\varphi \in C_b[0,\infty)$ *is given by*

$$\varphi(s) = \frac{1}{\sqrt{\pi}}\{\psi(s) + \int_0^\infty \psi'(s+v)[1-(1-e^{-2v})^{-1/2}]dv\}. \tag{5.8}$$

Proof. This is essentially calculation, the principal part of which we exhibit at the outset: for all $v > 0$

$$\begin{aligned} &\int_0^v \left(e^{2(v-s)}-1\right)^{-1/2}[1-(1-e^{-2s})^{1/2}]ds \\ &= \left[-Tan^{-1}((e^{2(v-s)}-1)^{1/2}) - \frac{1}{2}Sin^{-1}\left(\frac{2e^{2s}-e^{2v}-1}{e^{2v}-1}\right)\right]_0^v \\ &= Tan^{-1}((e^{2v}-1)^{1/2}) - \frac{\pi}{2}. \end{aligned} \tag{5.9}$$

Suppose then that $\psi \in C_b^1[0,\infty)$ and define ϕ by (5.8). $\phi \in C_b[0,\infty)$ since $[1-(1-e^{-2v})^{1/2}] \in L^1[0,\infty)$. Now a change of variable and interchange of order of integration show that

$$\begin{aligned}
&\frac{2}{\sqrt{\pi}}\int_0^\infty \left\{\int_0^\infty \psi'(s+v+u)[1-(1-e^{-2u})^{-1/2}]du\right\}(e^{2v}-1)^{-1/2}dv \\
&\qquad = \frac{2}{\sqrt{\pi}}\int_0^\infty \psi'(s+v)\int_0^v (e^{2(v-u)}-1)^{-1/2}[1-(1-e^{-2u})^{-1/2}]dudv \qquad (5.10)\\
&\qquad = \frac{2}{\sqrt{\pi}}\int_0^\infty \psi'(s+v)[Tan^{-1}((e^{2v}-1)^{1/2})-\frac{\pi}{2}]dv,
\end{aligned}$$

by the identity (5.9). Integrating by parts, this becomes

$$= \frac{2}{\sqrt{\pi}}\left[\frac{\pi}{2}\psi(s) - \int_0^\infty \psi(s+v)(e^{2s}-1)^{-1/2}ds\right].$$

After rearrangement this reduces to $\psi = \mathscr{B}[\varphi]$.

All that remains is to show that $\mathscr{B}$ is 1-1. First suppose only $\varphi' \in C_b[0,\infty)$, and let $\psi = \mathscr{B}[\varphi]$. Since φ grows at most linearly and $v\cdot(e^{2v}-1)^{-1/2} \in L^1[0,\infty)$, it follows that $\psi' \in C_b[0,\infty)$ and

$$\psi'(s) = \frac{2}{\sqrt{\pi}}\int_0^\infty \varphi'(s+v)(e^{2v}-1)^{-1/2}dv. \qquad (5.11)$$

Therefore

$$\begin{aligned}
&\int_0^\infty \psi'(s+u)[1-(1-e^{-2u})^{-1/2}]du \\
&\qquad = \frac{2}{\sqrt{\pi}}\int_0^\infty \left\{\int_0^\infty \varphi'(s+u+v)(e^{2v}-1)^{-1/2}dv\right\}[1-(1-e^{-2u})]^{-1/2}du.
\end{aligned}$$

Calculating as in (5.10), we see that

$$\begin{aligned}
\int_0^\infty \psi'(s+u)[1(1-e^{-2u})^{-1/2}]du &= \sqrt{\pi}\varphi(s) - \mathscr{B}[\varphi](s) \\
&= \sqrt{\pi}\varphi(s) - \psi(s).
\end{aligned}$$

That is, if $\varphi' \in C_b[0,\infty)$ then (5.8) recovers φ from $\psi = \mathscr{B}[\varphi]$. In particular, $\mathscr{B}$ is 1-1 on $\{\varphi \in C[0,\infty) : \varphi' \in C_b[0,\infty)\}$. To see that it must be 1-1 on all of $C_b[0,\infty)$, argue as follows. If $\varphi \in C_b[0,\infty)$ and $\mathscr{B}[\varphi] \equiv 0$, then consider

$$\Phi(t) = \int_0^t \varphi(u)du.$$

$\Phi'(t) = \varphi(t) \in C_b[0,\infty)$ and by (5.11)

$$\frac{d}{dt}\mathscr{B}[\Phi] = \mathscr{B}[\phi] \equiv 0$$

or $\mathscr{B}[\Phi] \equiv c$, some constant. By what we just proved, Φ is recovered by using $\psi \equiv c$ in (5.8):

$$\Phi(t) \equiv \frac{1}{\sqrt{\pi}}c.$$

Therefore $\varphi(t) \equiv \Phi'(t) \equiv 0$. □

Now we can collect the implications for $\overline{B}$.

THEOREM 5.1. *$\overline{B}$ is a 1-1 mapping of $C(\partial D)$ into itself. If $h = \overline{B}[g]$ then*

$$h(y) = \frac{2}{\sqrt{\pi}} \int_0^\infty g(x^0(s;y))(e^{2s} - 1)^{-1/2} ds. \tag{5.12}$$

If $h \in C^1(\partial D)$ then $h = \overline{B}[g]$, where $g \in C(\partial D)$ is given by

$$g(y) = \frac{1}{\sqrt{\pi}} \left\{ h(y) + \int_0^\infty \frac{d}{ds} h(x^0(s;y)) \cdot [1 - (1 - e^{-2s})^{-1/2}] ds \right\}. \tag{5.13}$$

In particular the range of $\overline{B}$ is dense in $C(\partial D)$.

Proof. If $g \in C(\partial D)$, then $h = \overline{B}[g]$ if and only if, for every $y \in \partial D$, $\psi = \mathscr{B}[\varphi]$, where $\psi(s) = h(x^0(s;y))$ and $\varphi(s) = g(x^0(s;y))$. Equation (5.12) follows from Lemma 5.2. If $h \equiv 0$ then $\psi \equiv 0$, so by the same lemma $\varphi \equiv 0$, i.e., $g = 0$ along every trajectory $x^0(s)$ on ∂D. Therefore $g \equiv 0$. That $h \in C(\partial D)$ follows from (5.12) and dominated convergence.

If $h \in C^1(\partial D)$ then $\frac{d}{ds} h(x^0(s;y))$ is bounded and continuous. Dominated convergence implies that g defined by (5.13) is in fact in $C(\partial D)$. That $h = \overline{B}[g]$ is Lemma 5.3 applied along every trajectory $x^0(s)$. The final assertion follows immediately since $C^1(\partial D)$ is dense in $C(\partial D)$. □

6. Inversion of the fundamental equation and a conjecture. We can now return to the consideration of our fundamental equation (3.7). Suppose $\{\epsilon_n\}_1^\infty$ is a sequence decreasing to 0 and that along this sequence $\nu^{\epsilon_n} \to \nu$ and $\mu^{\epsilon_n} \to \mu$ (weakly). Passing to the limit in (3.7) using Theorem 4.1, we obtain, for some constant c,

$$\int_{\partial D} f d\nu = c \int_{\partial D} B[f] d\mu = c \int_{\partial D} \overline{B}\left[\frac{f}{\gamma \alpha^2}\right] d\mu \quad \forall f \in C(\partial D).$$

This is equivalent to

$$\int_{\partial D} g \cdot \gamma \alpha^2 d\nu = c \int_{\partial D} \overline{B}[g] d\mu \quad \forall g \in C(\partial D). \tag{6.1}$$

Since the range of $\overline{B}$ is dense in $C(\partial D)$, this equation determines μ uniquely in terms of ν. To be explicit, using (5.13) in (6.1) we have that for all $h \in C^1(\partial D)$,

$$\int_{\partial D} h d\mu = c \cdot \left\{ \int_{\partial D} h \gamma \alpha^2 d\nu + \int_0^\infty \left(\int_{\partial D} \frac{d}{ds} h(x^0(s;y)) \gamma(y) \alpha(y)^2 d\nu \right) \cdot [1 - (1 - e^{2s})^{-1/2}] ds \right\}, \tag{6.2}$$

where again c is the appropriate normalizing constant. It might be natural to write

$$\frac{d}{ds} h(x^0(s;y)) = \langle \nabla h(x^0), \bar{b}(x^0) \rangle, \quad \text{where } x^0 = x^0(s;y),$$

although technically since h is only defined in ∂D, only the tangential components of ∇h are defined. The point is simply that $(d/ds)h(x^0)$ is a first-order differential operator applied to h evaluated at $x^0(s) \in \partial D$.

Our conclusions for the exit problem are collected in the following theorem.

THEOREM 6.1. *Let $\nu^\epsilon(dy)$ be the probability measures on ∂D defined by*

$$\nu^\epsilon(dy) = c^\epsilon p^\epsilon(y) dy.$$

For any $x \in D$ and sequence $\{\epsilon_n\}_1^\infty$ decreasing to 0, the exit measures $\mu_x^{\epsilon_n}(dy)$ converge (weakly) to a measure $\mu(dy)$ if and only if $\nu^{\epsilon_n}(dy)$ converges (weakly) to a measure $\nu(dy)$, in which case $\mu(dy)$ is uniquely determined by $\nu(dy)$ according to (6.1) or (6.2).

Proof. Everything has already been argued except the "if and only if" assertion. If μ^{ϵ_n} converges to a probability measure μ, and $f \in C(\partial D)$ then we know from (3.7) that

$$\int_{\partial D} f d\nu^\epsilon = c^\epsilon \int_{\partial D} B^\epsilon[f] d\mu^\epsilon = \epsilon^{-1/2} c^\epsilon \cdot \int_{\partial D} \epsilon^{1/2} B^\epsilon[f] d\mu^\epsilon \tag{6.3}$$

and

$$\int_{\partial D} \epsilon^{1/2} B^\epsilon[f] d\mu^\epsilon \to \int_{\partial D} B[f] d\mu.$$

First consider $f \equiv 1$. Since the left side in (6.3) is 1, we conclude that $\epsilon^{-1/2} c^\epsilon$ converges to the constant c which satisfies

$$1 = c \int_{\partial D} B[1] d\mu.$$

Thus for an arbitrary $f \in C(\partial D)$ we have

$$\int_{\partial D} f d\nu^{\epsilon_n} \to c \int_{\partial D} B[f] d\mu.$$

The right side of this defines a probability measure ν on ∂D. This proves the "only if" assertion.

Suppose ν^{ϵ_n} converges to ν. Since ∂D is compact, the set of probability measures on ∂D is precompact. All convergent subsequences of $\{\mu^{\epsilon_n}\}$ must have the same limit μ, uniquely determined from ν by μ. Thus $\{\mu^{\epsilon_n}\}$ converges. This proves the "if". □

Perhaps it is disheartening that the "inversion formula" (6.2) is not more simple (compare (6.6) of [2] for the case of (1.5)). However (6.2) is the best that can be said without more information about ν. (It is tempting to integrate by parts in (6.2), however $(d/ds)[1-(1-e^{-2s})^{-1/2}] \notin L^1$.) As indicated in the introduction, we are not going to pursue the asymptotic analysis of p^ϵ and implications for ν here. However, we will make the following conjecture:

CONJECTURE. *If $\nu^\epsilon \to \nu$ weakly as $\epsilon \downarrow 0$, then $\gamma(y)\alpha(y)^2\nu(dy)$ is necessarily an invariant measure for the boundary flow $x^0(s)$ with respect to the variable s : $\dot{x}^0(s) = \bar{b}(x^0(s))$. That is*

$$\int_{\partial D} f(y)\gamma(y)\alpha(y)^2\nu(dy) = \int_{\partial D} f(x^0(s;y))\gamma(y)\alpha(y)^2\nu(dy)$$

for all $s \geq 0, f \in C(\partial D)$.

This conjecture agrees with the formal calculations of [11]. The following theorem allows an equivalent formulation of the conjecture in terms of $\mu = \lim \mu^\epsilon$.

THEOREM 6.2. *Let ν and μ be the respective limits of sequences $\{\nu^{\epsilon_n}\}$ and $\{\mu^{\epsilon_n}\}$, as in Theorem* 6.1. *The measure $\gamma(y)\alpha(y)^2\nu(dy)$ is invariant for $x^0(s;y)$ if and only if $\mu(dy)$ is also, in which case*

$$\mu(dy) = c\gamma(y)\alpha(y)^2\nu(dy), \tag{6.4}$$

with the appropriate normalizing constant c.

Proof. If $\gamma(y)\alpha(y)^2\nu(dy)$ is invariant as claimed, then for $h \in C^1(\partial D)$ for all $s \geq 0$,

$$0 = \frac{d}{ds}\int_{\partial D} h(x^0(s;y))\gamma(y)\alpha(y)^2\nu(dy) = \int_{\partial D} \frac{d}{ds} h(x^0(s;y))\gamma(y)\alpha(y)^2\nu(dy).$$

With this fact (6.2) reduces to (6.4) and of course μ is therefore $x^0(s)$ invariant.

Conversely, suppose μ is $x^0(s)$ invariant. Then using (6.1) and (5.12),

$$\begin{aligned}\int_{\partial D} g(x^0(s;y))\gamma(y)\alpha(y)^2\nu(dy) &= c\int_{\partial D}\overline{B}[g(x^0(s;\cdot))]\mu(dy)\\ &= c\int_0^\infty \frac{2}{\sqrt{\pi}}(e^{2v}-1)^{-1/2}\left[\int_{\partial D} g(x^0(s+v;y))\mu(dy)\right]ds\\ &= c\int_0^\infty \frac{2}{\sqrt{\pi}}(e^{2v}-1)^{-1/2}\left[\int_{\partial D} g(x^0(s;y))\mu(dy)\right]ds\\ &= c\int_{\partial D}\overline{B}[g]\mu(dy)\\ &= \int_{\partial D} g(y)\gamma(y)\alpha(y)^2\nu(dy).\end{aligned}$$

This establishes the invariance of $\gamma(y)\alpha(y)^2\nu(dy)$. □

To close this section consider the case studied in [11]: $d=2$ and ∂D consisting of a single periodic orbit of (1.4), $|b(y)|>0$ everywhere on ∂D. There is then a unique (up to a scalar) $x^0(s)$ invariant measure on ∂D given by

$$\frac{dy}{|\bar{b}(y)|} = \frac{\alpha(y)^2\gamma(y)^2}{|b(y)|}dy.$$

Thus, supposing our conjecture above to be true, there is only one possible limit ν of ν^ϵ as $\epsilon \downarrow 0$. Therefore if $\mu^\epsilon(dy)$ converges, then its limit must be the measure

$$\mu(dy) = c\frac{\alpha(y)^2\gamma(y)^2}{|b(y)|}.$$

This agrees with (2.46) of [11] since $Z(y)=\gamma(y)^{-2}$ and $\alpha(y)\equiv 1$ (they assumed $a(x)\equiv I$).

7. Some estimates. The remainder of this paper is technical, leading to proofs of Theorems 2.1 and 4.1. In this section we will derive upper bounds for several quantities associated with the processes $\varsigma^\epsilon(t)$ and $\varsigma(t)$, as defined in §4. In each case the idea is the same: calculate the analogous quantity explicitly for a one-dimensional diffusion in $[0,\infty]$ with generator

$$Gf(z) = \frac{1}{2}\alpha_0^2 f''(z) + k(z)f'(z)$$

and instantaneous reflection at 0. Then use Itô's lemma to show that this quantity bounds the desired quantity for ς^ϵ and ς. The key to success is a careful choice of the drift coefficient $k(z)$.

We will take

$$k(z) = \frac{\alpha_0^2}{A_1}Kz - 1,$$

where K is the positive constant of (2.11). Note then that for any $x\in G$, with $\varsigma = \epsilon^{-1/2}\rho(x)$,

$$\begin{aligned}\epsilon^{-1/2}\mathscr{L}^\epsilon\rho(x) &= \epsilon^{-1/2}\langle -n(x), b(x)\rangle + \frac{\epsilon^{1/2}}{2}\sum_{i,j} a_{ij}\rho_{x_i,x_j}\\ &\geq K\cdot\epsilon^{-1/2}\rho(x) + \epsilon^{1/2}A_0\\ &\geq K\varsigma - \epsilon^{1/2}|A_0|\end{aligned}$$

where A_0 is as in (2.5). Using (2.6) and (2.7) we have that

$$\begin{aligned}\left(\frac{\alpha_0}{\alpha(x)}\right)^2 \epsilon^{-1/2}\mathscr{L}^\epsilon \rho(x) &\geq \left(\frac{\alpha_0}{\alpha(x)}\right)^2 K\varsigma - \left(\frac{\alpha_0}{\alpha(x)}\right)^2 \epsilon^{1/2}|A_0| \\ &\geq \frac{\alpha_0^2}{A_1} K\varsigma - \epsilon^{1/2}|A_0| \\ &\geq \frac{\alpha_0^2}{A_1} K\varsigma - 1 \\ &= k(\varsigma),\end{aligned} \tag{7.1}$$

provided $\epsilon^{1/2}|A_0| \leq 1$. Define $\epsilon_0 = |A_0|^{-1/2}$, or $\epsilon_0 = 1$ if $A_0 = 0$. Then (7.1) holds for all $0 < \epsilon \leq \epsilon_0$. This value of ϵ_0 will remain unchanged in all that follows. Note also that (2.11) implies that $b_1(y) \geq K$ everywhere on ∂D. Therefore for all $y \in \partial D, \varsigma \geq 0$

$$\left(\frac{\alpha_0}{\alpha(y)}\right)^2 b_1(y)\varsigma \geq \frac{\alpha_0^2}{A_1} K\varsigma \geq k(\varsigma). \tag{7.2}$$

The following lemma contains our basic comparison argument. The stopping times η_r^ϵ were defined in (4.3). Their analogues for $\varsigma(t)$ are

$$\eta_r = \inf\{t > 0 : \varsigma(t) = r\},$$

defined for $0 < \varsigma(0) < r$.

LEMMA 7.1. *Suppose* $0 < r < \infty$ *and* $u \in C^2([0,r])$ *is nonincreasing. Suppose that for some constant* $c \geq 0$

$$Gu(z) + c = 0 \quad \forall\, 0 \leq z \leq r.$$

Then for any stopping time $\tau \leq \eta_r^\epsilon$, *all* $0 < \epsilon \leq \epsilon_0$ *and* $x_0 \in G$ *with* $\varsigma^\epsilon(0) < r \leq \epsilon^{-1/2}h_0$, *the following holds:*

$$E_{x_0}[u(\varsigma^\epsilon(\tau))] + cE_{x_0}[\tau] - \frac{1}{2}\alpha_0^2 u'(0)E_{x_0}[\epsilon^{1/2}l^\epsilon(\tau)] \leq u(\varsigma^\epsilon(0)) \leq u(0) \tag{7.3}$$

Likewise, in the case of $\varsigma(t)$, *for any stopping time* $\tau \leq \eta_r, 0 \leq \varsigma(0) < r, x^0(0) = y \in \partial D$, *the following holds*

$$E_{\varsigma(0),y}[u(\varsigma(\tau)); \tau < \infty] + cE_{\varsigma(0),y}[\tau] - \frac{1}{2}\alpha_0^2 u'(0)E_{\varsigma(0),y}[l(\tau)] \leq u(\varsigma(0)) \leq u(0). \tag{7.4}$$

We also allow $r = \infty, \tau \leq \infty$ *in the case of* $\varsigma(t)$, *provided* u *is bounded.*

Proof. We will give the argument for $\varsigma^\epsilon(t)$, the case of $\varsigma(t)$ is analogous. Applying Itô's formula,

$$du(\varsigma^\epsilon(t)) = \mathscr{G}^\epsilon u(t)dt + \alpha(x^\epsilon(t))u'(\varsigma^\epsilon(t))d\beta(t) + \frac{\epsilon^{1/2}}{2}u'(0)\alpha(x^\epsilon(t))^2 dl^\epsilon(t),$$

where

$$\begin{aligned}\mathscr{G}^\epsilon u(t) &= \frac{1}{2}\alpha(x^\epsilon(t))^2 u''(\varsigma^\epsilon(t)) + [\epsilon^{-1/2}\mathscr{L}^\epsilon \rho(x^\epsilon(t))] \cdot u'(\varsigma^\epsilon(t)) \\ &\leq \frac{\alpha(x^\epsilon(t))^2}{\alpha_0^2}\left[\frac{1}{2}\alpha_0^2 u''(\varsigma^\epsilon) + k(\varsigma^\epsilon)u'(\varsigma^\epsilon)\right] \leq -c.\end{aligned}$$

The first inequality follows from (7.1) and the hypothesis that $u' \leq 0$; the second follows from (2.7) and $c \geq 0$. Thus for any finite T,

$$E[u(\varsigma^\epsilon(T \wedge \tau))] - u(\varsigma^\epsilon(0)) \leq -cE[T \wedge \tau] + \frac{1}{2}u'(0)\alpha_0^2 E[\epsilon^{1/2}l^\epsilon(T \wedge \tau)]. \tag{7.5}$$

Rearranging (7.5), and using the monotonicity of u,

$$E[u(\zeta^\epsilon(T\wedge\tau))]+cE[T\wedge\tau]-\frac{1}{2}\alpha_0^2u'(0)E[\epsilon^{1/2}l^\epsilon(T\wedge\tau)]\le u(\zeta^\epsilon(0))\le u(0)$$

Letting $T\to\infty$ yields (7.3). □

We now exhibit the particular choices of u that yield the esimates we want. Define

$$\Phi(z)=\frac{2}{\alpha_0^2}\int_0^z k(s)ds \text{ for } z\ge 0.$$

It follows from $K>0$ that $e^{-\Phi(z)}$ is integrable. Take

$$u_1(z)=\frac{2}{\alpha_0^2}\int_z^\infty e^{-\Phi(s)}ds \quad \textit{for all } z\ge 0,$$

$$u_2(z)=\frac{u_1(z)}{u_1(0)},$$

and, for any specified $r>0$,

$$u_{3,r}(z)=2\alpha_0^{-2}\int_z^r\int_0^s e^{\Phi(u)-\Phi(s)}duds \text{ for } 0\le z\le r.$$

Each of these is C^2 and nonincreasing.

In the case of u_1 we have

$$Gu_1=0, u_1'(0)=-\frac{2}{\alpha_0^2}.$$

LEMMA 7.2. *For all* $0<\epsilon\le\epsilon_0, x_0\in G$ *with* $\zeta^\epsilon(0)<\epsilon^{-1/2}h_0, y\in\partial D$ *and* $\zeta_0>0$,

$$E_{x_0}[\epsilon^{1/2}l^\epsilon(\tau_\Gamma)]\le u_1(0)$$
$$E_{\zeta_0,y}[l(\infty)]\le u_1(0).$$

Proof. The first follows by using $r=\epsilon^{-1/2}h_0$ in Lemma 7.1, noting that $\eta_r^\epsilon=\tau_\Gamma$, and using $u_1(r)\ge 0$. The second uses $r=\infty$ and $u_1(\infty)=0$. □

Considering u_2 yields the next estimate.

LEMMA 7.3. *For all* $0<\epsilon\le\epsilon_0, x_0\in G$ *with* $\zeta^\epsilon(0)<\epsilon^{-1/2}h_0, y\in\partial D$ *and* $\zeta_0>0$,

$$P_{x_0}[\tau_{\partial D}<\tau_\Gamma]\le u_2(\zeta^\epsilon(0))$$
$$P_{\zeta_0,y}[\eta_0<\infty]\le u_2(\zeta_0)$$

Proof. Again $Gu_2=0$. For the case of ζ^ϵ let $r=\epsilon^{-1/2}h_0$ and $\tau=\tau_{\partial D}\wedge\eta_r^\epsilon=\tau_{\partial D}\wedge\tau_\Gamma$. Since $u_2'(0)\le 0, u_2(0)=1$, and $u_2(r)\ge 0$

$$\begin{aligned}P_{x_0}[\tau_{\partial D}<\tau_\Gamma]&=E_{x_0}[u_2(\zeta^\epsilon(\tau));\tau=\tau_{\partial D}]\\&\le E_{x_0}[u_2(\zeta^\epsilon(\tau))]\\&\le u_2(\zeta^\epsilon(0)).\end{aligned}$$

The argument is similar for ζ, using $r=\infty$ and $\tau=\eta_0$. □

Using $u_{3,r}$, we obtain the following.

LEMMA 7.4. *For all* $0 < \epsilon \le \epsilon_0, x_0 \in G$ *with* $\varsigma^\epsilon(0) < r \le \epsilon^{-1/2}h_0, y \in \partial D$ *and* $\varsigma_0 < r$

$$E_{x_0}[\eta_r^\epsilon] \le u_{3,r}(0)$$
$$E_{zz_0,y}[\eta_r] \le u_{3,r}(0).$$

Proof. This time we have

$$Gu_{3,r} = -1;\ u'_{3,r}(0) = 0,\ u_{3,r}(r) = 0.$$

Simply apply Lemma 7.1 using $\tau = \eta_r^\epsilon$ (or η_r in the case of $\varsigma(t)$) and observe that $0 \le E[u(\varsigma^\epsilon(\tau))]$.

COROLLARY 7.1. *For all* $0 < \epsilon \le \epsilon_0, x_0 \in G$ *with* $\varsigma^\epsilon(0) < r \le \epsilon^{1/2}h_0, y \in \partial D, \varsigma_0 < r$ *and* $T < \infty$,

$$P_{x_0}[\eta_r^\epsilon > T] \le \frac{1}{T}u_{3,r}(0)$$
$$P_{\varsigma_0,y}[\eta_r > T] \le \frac{1}{T}u_{3,r}(0)$$

Proof. The proof is Chebyshev's inequality. □

8. Proof of Theorem 2.1. Take any fixed $\epsilon > 0$. To prove Theorem 2.1 consider the process $r(t) = \rho(x^\epsilon(t))$ defined on $[0, \tau_\Gamma]$. According to Itô's formula

$$dr(t) = \mathscr{L}^\epsilon\rho(x^\epsilon(t))dt + \epsilon^{1/2}\langle\sigma^T\nabla\rho(x^\epsilon(t)), dw(t)\rangle + \frac{\epsilon}{2}\alpha(x^\epsilon(t))^2 dl^\epsilon(t),$$

since $\langle\nabla\rho(y), \eta(y)\rangle = -\alpha(y)^2$ for $y \in \partial D$. For each $h > 0$ define a C^1, piecewise C^2 function $\psi_h : [0, \infty) \to \mathbb{R}$ by

$$\psi_h''(u) = -h^{-1}\chi_{[0,h]}(u),\ \psi_h'(0) = 1,\ \psi_h(0) = 0.$$

It is elementary to check that

$$\psi_h'(u) = \begin{cases} 1 - u/h & 0 \le u \le h \\ 0 & h \le u \end{cases} \tag{8.1}$$

and

$$0 \le \psi_h(u) \le \frac{1}{2}h. \tag{8.2}$$

We want to apply Itô's formula to $\psi_h(r(t)) \cdot f(x^\epsilon(t))$, so assume that $f \in C^2(G)$. (Although ψ_h is not C^2, we can use a sequence of smooth approximants to justify the following.)

$$\begin{aligned} d[\psi_h(r(t))f(x^\epsilon(t))] = {} & [f \cdot \psi_h' \cdot \mathscr{L}^\epsilon\rho + \frac{\epsilon}{2}f\psi_h''\langle\nabla\rho, a\nabla\rho\rangle]dt \\ & + f \cdot \psi_h' \cdot \epsilon^{1/2}\langle\sigma^T\nabla\rho, dw\rangle + \frac{\epsilon}{2}f\psi_h'\alpha^2 dl^\epsilon \\ & + \psi_h \cdot \mathscr{L}^\epsilon f dt + \psi_h \cdot \epsilon^{1/2}\langle\sigma^T\nabla f, dw\rangle - \frac{\epsilon}{2}\psi_h\langle\nabla f, \eta\rangle dl^\epsilon \\ & + \frac{\epsilon}{2}\sum_1^n \psi_h' \cdot \langle\nabla\rho, a\nabla f\rangle dt. \end{aligned}$$

By construction, on ∂D we know that $\psi_h' = 1$ and $\psi_h \langle \nabla f, \eta \rangle = 0$. Because dl^ϵ is supported on $\{t : x^\epsilon(t) \in \partial D\}$, we can express the above as

$$\begin{aligned}\int_0^\tau \frac{1}{h}\chi_{[0,h]}(r) f(x) \cdot \langle n, an\rangle dt = &\int_0^{\tau_\Gamma} f(x^\epsilon)\alpha(x^\epsilon)^2 dl^\epsilon(t) \\ &+ \int_0^{\tau_\Gamma} \psi_h \cdot \Phi_1 dt + \int_0^{\tau_\Gamma} \psi_h' \cdot \Phi_2 dt \\ &+ \int_0^{\tau_\Gamma} \psi_h \cdot \langle \Phi_3, dw\rangle + \int_0^{\tau} \psi_h' \langle \Phi_4, dw\rangle \\ &+ f(x^\epsilon(0))\psi_h(r(0)) - f(x^\epsilon(\tau_\Gamma))\psi_h(r(\tau_\Gamma)).\end{aligned} \tag{8.3}$$

Here $\Phi_i, i = 1, \cdots, 4$ are bounded continuous functions on G (depending on f and ϵ) evaluated at $x^\epsilon(t)$.

It follows from Lemma 7.4 that for $\epsilon > 0$ fixed, $E_{x_0}[\tau_\Gamma]$ is uniformly bounded over $x_0 \in G$. This together with (8.2) implies that all the terms in (8.3) with a factor of ψ_h converge to 0 in the mean, as $h \downarrow 0$. From (8.1) it follows that the terms involving ψ_h' converge in the mean to

$$\int_0^{\tau_\Gamma} \chi_{\{0\}}(r(t))\Phi_2 dt$$

and

$$\int_0^{\tau_\Gamma} \chi_{\{0\}}(r(t))\langle \Phi_4, dw\rangle.$$

But these are both 0 almost surely, by virtue of (2.15). This proves Theorem 1 for $f \in C^2(G)$. The general $f \in C(G)$ can be uniformly approximated by $C^2(G)$ functions to yield Theorem 1 in its stated generality.

9. Proof of Theorem 4.1. Our proof of Theorem 4.1 will use the fact that, on bounded time intervals, $x^\epsilon(t)$ converges to $x^0(t)$ uniformly in t, and uniformly over all $x^\epsilon(0) = x^0(0) = y \in \partial D$. This follows from large deviations results, such as in [1]. However, it is not necessary to appeal to such deep results. We will give an independent proof, suited to our particular context.

Two special notations will be useful. First, for $\phi : [0, T] \to \mathbb{R}^d$ (or $\mathbb{R}^1$), define

$$[\phi]_t = \sup_{0 \le u \le t} |\varphi(u)|, t \le T.$$

Secondly, for functions $f : \partial D \to \mathbb{R}$ define

$$\|f\|_{\partial D} = \sup_{y \in \partial D} |f(y)|.$$

At several stages below we will have a one-dimensional Skorohod equation of the form

$$\psi(t) = \varphi(t) + m(t); \psi(t) \ge 0, \tag{9.1}$$

where all 3 functions are continuous and $m(t)$ is nondecreasing and obeys

$$dm(t) = \chi_{\{0\}}(\psi(t))dm(t).$$

We will use the fact that the solution is determined uniquely from $\varphi(\cdot)$ by the formula

$$m(t) = -\inf_{0 \le u \le t}(\varphi(u) \wedge 0); \tag{9.2}$$

see [9, Lemma 4.2, p. 119].

LEMMA 9.1. $[x^\epsilon - x^0]_T \to 0$ *in probability, uniformly over* $x^\epsilon(0) = x^0(0) = y \in \partial D$; *i.e., given any* $T, \delta > 0$

$$\sup_{y \in D} P_y[[x^\epsilon - x^0]_T > \delta] \to 0 \text{ as } \epsilon \downarrow 0.$$

Proof. It suffices to consider $0 < \delta < h_0$. Since $x^0(0) = y \in \partial D$ we know that $x^0(t) \in \partial D$, for all $t \geq 0$. Thus

$$\{\tau_\Gamma < T\} \subseteq \{[x^\epsilon - x^0]_{\tau_\Gamma \wedge T} > \delta\}.$$

Therefore

$$\{[x^\epsilon - x^0]_T > \delta\} \subseteq \{[x^\epsilon - x^0]_T > \delta; \tau_\Gamma \geq T\} \cup \{\tau_\Gamma < T\}$$

$$\subseteq \{[x^\epsilon - x^0]_{\tau_\Gamma \wedge T} > \delta\}.$$

Now applying Itô's formula, we have, for $0 \leq t \leq \tau_\Gamma$,

$$\rho(x^\epsilon(t)) = \Phi^\epsilon(t) + \frac{\epsilon}{2}\alpha(x^\epsilon(t))^2 l^\epsilon(t), \ \rho(x^\epsilon(t)) \geq 0$$

$$l^\epsilon(t) = \int_0^t \chi_{\{0\}}(\rho(x^\epsilon(t))dl^\epsilon(t),$$

where

$$\Phi^\epsilon(t) = \int_0^t \mathscr{L}^\epsilon \rho(x^\epsilon(s))ds + \epsilon^{1/2}\int_0^t \langle \sigma^T \nabla \rho(x^\epsilon(s)), dw(s)\rangle.$$

This is a one-dimensional Skorohod equation (9.1), so that by (9.2) we have (almost surely)

$$\frac{\epsilon}{2}\alpha_0^2 l^\epsilon(t) \leq \frac{\epsilon}{2}\alpha(x^\epsilon(t))^2 l^\epsilon(t) = -\inf_{[0,t]}(\Phi^\epsilon(t) \wedge 0).$$

Now by (2.5) and (2.12),

$$\Phi^\epsilon(t) = \int_0^t \frac{\epsilon}{2}\sum a_{ij}\rho_{x_i x_j}(x^\epsilon(s)) + \langle b(x^\epsilon(s)), \nabla\rho(x^\epsilon(s))\rangle ds$$

$$+ \epsilon^{1/2}\int_0^t \langle \sigma^T \nabla \rho(x^\epsilon(s)), dw(s)\rangle$$

$$\geq \epsilon A_0 t - \epsilon^{1/2}\left|\int_0^t \langle \sigma^T \nabla \rho, dw(s)\rangle\right|$$

Thus for all $0 \leq t \leq \tau_\Gamma \wedge T$,

$$\frac{\epsilon}{2}\alpha_0^2 l^\epsilon(t) \leq \epsilon|A_0|t + \epsilon^{1/2}\left[\int_0^\cdot \langle \sigma^T \nabla \rho(x^\epsilon(s)), dw(s)\rangle\right]_{T \wedge \tau_\Gamma}$$

But a standard application of Gronwall's inequality applied to (2.13) tells us that

$$[x^\epsilon - x^0]_{\tau_\Gamma \wedge T} \leq e^{MT}\left[\epsilon^{1/2}\int_0^\cdot \sigma dw(s) - \int_0^\cdot \frac{\epsilon}{2}\eta(x^\epsilon)dl^\epsilon\right]_{\tau_\Gamma \wedge T}$$

$$\leq e^{MT}\left(\epsilon^{1/2}\left[\int_0^\cdot \sigma dw\right]_{\tau_\Gamma \wedge T} + \frac{\epsilon}{2}C[l^\epsilon]_{\tau_\Gamma \wedge T}\right),$$

where M is a Lipschitz constant for $b(\cdot)$, and C is a bound on $|\eta(y)|$ over ∂D. Therefore

$$
\begin{aligned}
[x^\epsilon - x^0]_{\tau_\Gamma \wedge T} \leq e^{MT} \Bigg(& \epsilon |A_0| \frac{C}{\alpha_0^2} T + \epsilon^{1/2} \left[\int_0^{\cdot} \sigma(x^\epsilon(s)) dw(s) \right]_{\tau_\Gamma \wedge T} \\
& + \epsilon^{1/2} \frac{C}{\alpha_0^2} \left[\int_0^{\cdot} \langle \sigma^T \nabla \rho(x^\epsilon(s)), dw(s) \rangle \right]_{\tau_\Gamma \wedge T} \Bigg).
\end{aligned} \tag{9.3}
$$

By a standard estimation argument (e.g. [15, (2.1), p. 87]),

$$
\begin{aligned}
& P_y \left[\left[\int_0^{\cdot} \sigma(x^\epsilon(s)) dw(s) \right]_T \geq \lambda \right]_T \leq 2d \cdot \exp\left(\frac{-\lambda^2}{2dA_1 T} \right) \\
& P_y \left[\left[\int_0^{\cdot} \langle \sigma \nabla \rho(x^\epsilon(s)), dw(s) \rangle \right]_{T \wedge \tau_\Gamma} \geq \lambda \right] \leq 2 \exp\left(\frac{-\lambda^2}{2A_1 T} \right),
\end{aligned} \tag{9.4}
$$

where A_1 is as in (2.6). Using these estimates in (9.3) implies the lemma. □

Next we want to establish convergence of $\varsigma^\epsilon(\cdot)$, defined by (4.1), to $\varsigma(\cdot)$ of (4.7), and $\epsilon^{1/2} l^\epsilon(\cdot)$ to $l(\cdot)$. Defining the one-dimensional Brownian motion $\beta(\cdot)$ by (4.6), (4.4) becomes

$$
d\varsigma^\epsilon(t) = \epsilon^{1/2} \mathscr{L}^\epsilon \rho(x^\epsilon(t)) + \alpha(x^\epsilon(t)) d\beta(t) + \frac{\epsilon^{1/2}}{2} dl^\epsilon(t).
$$

We noted previously that the $\beta(\cdot)$ here is ϵ-dependent and distinct from the $\beta(\cdot)$ in (4.7), so we should talk about weak or distributional convergence of $\varsigma^\epsilon(\cdot)$ to $\varsigma(\cdot)$. However, the distribution of $\varsigma(\cdot)$ is the same regardless of the choice of $\beta(\cdot)$ in (4.7). So, for each ϵ, we can take the $\beta(\cdot)$ of (4.7) to be that defined by (4.6), with a suitable extension to $t > \tau_\Gamma$. Now $\varsigma^\epsilon(\cdot)$ and $\varsigma(\cdot)$ are defined on a common (but ϵ-dependent) probability space so we can make pathwise comparisons. We do, however, need to limit $x^\epsilon(0) = x^0(0) = y$ to points $y \in \partial D$ and $\varsigma(0) = \varsigma^\epsilon(0) = 0$, since (4.7) is only well defined for $x^0(\cdot) \in \partial D$. Under these conventions, the expectation $E_{0,y}$ (associated with $\varsigma(\cdot)$) coincides with E_y (associated with $x^\epsilon(\cdot)$ and $\varsigma^\epsilon(\cdot)$); i.e., both symbols refer to integrals with respect to the same probability measure on the same probability space. With this understanding we have the following result.

LEMMA 9.2. *$[\varsigma^\epsilon - \varsigma]_{T \wedge \eta_r^\epsilon}$ and $[\epsilon^{1/2} l^\epsilon - l]_{T \wedge \eta_r^\epsilon}$ converge to 0 in L^2, uniformly over $y = x^\epsilon(0) = x^0(0) \in \partial D$; i.e., given any $T, \delta, r > 0$,*

$$
\sup_{y \in \partial D} E_y [[\varsigma^\epsilon - \varsigma]^2_{T \wedge \eta_r^\epsilon}] \to 0,
$$

and

$$
\sup_{y \in \partial D} E_y [[\epsilon^{1/2} l^\epsilon - l]^2_{T \wedge \eta_r^\epsilon}] \to 0,
$$

as $\epsilon \downarrow 0$.

Proof. Let us abbreviate (4.4) and (4.7) as

$$
\begin{aligned}
& \varsigma^\epsilon(t) = \Psi^\epsilon(t) + M^\epsilon(t), \ M^\epsilon(t) = \int_0^t \chi_{\{0\}}(\varsigma^\epsilon(s)) dM^\epsilon(s) \\
& \varsigma(t) = \Psi(t) + M(t), \ M(t) = \int_0^t \chi_{\{0\}}(\varsigma(s)) dM(s).
\end{aligned} \tag{9.5}
$$

where

$$
\Psi^\epsilon(t) = \int_0^t \epsilon^{-1/2} \mathscr{L}^\epsilon \rho(x^\epsilon(s)) ds + \int_0^t \alpha(x^\epsilon(s)) d\beta(s),
$$

$$\Psi(t) = \int_0^t b_1(x^0(s))\varsigma(s)ds + \int_0^t \alpha(x^0(s))d\beta(s),$$

and

$$M^\epsilon(t) = \frac{\epsilon^{1/2}}{2}\int_0^t \alpha(x^\epsilon(s))^2 dl^\epsilon(s)$$

$$M(t) = \frac{1}{2}\int_0^t \alpha(x^0(s))^2 dl(s).$$

From (9.2) we have

$$M^\epsilon(t) = -\inf_{[0,t]}(0 \wedge \Psi^\epsilon),\ M(t) = -\inf_{[0,t]}(0 \wedge \Psi).$$

Therefore for $t \leq \tau_\Gamma$

$$[M^\epsilon - M]_t \leq [\Psi^\epsilon - \Psi]_t, \tag{9.6}$$

and, by subracting the two equations in (9.5),

$$[\varsigma^\epsilon - \varsigma]_{t\wedge\eta_r^\epsilon} \leq 2[\Psi^\epsilon - \Psi]_{t\wedge\eta_r^\epsilon}. \tag{9.7}$$

Now we also know that

$$\Psi^\epsilon(t) - \Psi(t) = \int_0^t b_1(x^\epsilon(s))\varsigma^\epsilon(s) - b_1(x^0(s))\varsigma(s) + o(1)ds + \int_0^t \alpha(x^\epsilon(s)) - \alpha(x^0(s))d\beta(s),$$

and so

$$[\Psi^\epsilon - \Psi]_t \leq \int_0^t |b_1(x^0(s))| \cdot [\varsigma^\epsilon - \varsigma]_s + [b_1(x^\epsilon(\cdot)) - b_1(x^0(\cdot))]_s|\varsigma^\epsilon(s)| + o(1)\ ds + \left[\int_0^\cdot \alpha(x^\epsilon(s)) - \alpha(x^0(s))d\beta(s)\right]_t. \tag{9.8}$$

Recall that the $o(1)$ is uniform so long as ς^ϵ stays in a compact set. Since $|\varsigma^\epsilon(s)| \leq r$ for $s \leq \eta_r^\epsilon$, we can write

$$[\Psi^\epsilon - \Psi]_{t\wedge\eta_r^\epsilon} \leq \int_0^t N[\varsigma^\epsilon - \varsigma]_{s\wedge\eta_r^\epsilon}ds + D_b^\epsilon + D_\beta^\epsilon + o(1) \tag{9.9}$$

where

$$N = \sup_{y\in\partial D}|b_1(y)|$$

$$D_b^\epsilon = T \cdot r \cdot [b_1(x^\epsilon) - b_1(x^0)]_T$$

$$D_\beta^\epsilon = \left[\int_0^\cdot \alpha(x^\epsilon(s)) - \alpha(x^0(s))d\beta\right]_{T\wedge\eta_r^\epsilon}.$$

Putting (9.7) and (9.9) together, we can apply Gronwall's lemma to conclude that

$$[\varsigma^\epsilon - \varsigma]_{T\wedge\eta_r^\epsilon} \leq 2[o(1) + D_b^\epsilon + D_\beta^\epsilon]e^{2NT}.$$

We want to argue that $\|D_b^\epsilon\|_2$ and $\|D_\beta^\epsilon\|_2$ converge to 0 as $\epsilon \downarrow 0$, uniformly over $y \in \partial D$. For D_b^ϵ this follows from Lemma 9.1 and the continuity of b_1 on G. For D_β^ϵ,

we can use Doob's martingale inequality to estimate

$$P[|D_\beta^\epsilon| \geq \lambda] \leq \lambda^{-2} E\left[\left(\int_0^{T\wedge\eta_r^\epsilon} \alpha(x^\epsilon(s)) - \alpha(x^0(s))d\beta(s)\right)^2\right]$$
$$\leq \lambda^{-2} E\left[\int_0^T |\alpha(x^\epsilon(s)) - \alpha(x^0(s))|^2 ds\right],$$

which converges to 0, uniformly over $y \in \partial D$, by virtue of Lemma 9.1 and the continuity of α. Thus $D_\beta^\epsilon \to 0$ in probability, uniformly over ∂D. Next since $\|D_\beta^\epsilon\|_p$ is bounded in y, any $2 < p < \infty$ (by virtue of an estimate similar to (9.4)), $|D_\beta^\epsilon|^2$ is uniformly integrable over $\epsilon > 0$ and $y \in \partial D$. These two facts together imply that $\|D_\beta^\epsilon\|_2 \to 0$ as claimed.

We have now shown that $[\varsigma^\epsilon - \varsigma]_{T\wedge\eta_r^\epsilon}$ and, by (9.9), $[\Psi^\epsilon - \Psi]_{T\wedge\eta_r^\epsilon}$ both $\to 0$ in L^2. It follows from (9.5) that $[M^\epsilon - M]_{T\wedge\eta_r^\epsilon} \to 0$ in L^2 as well. To finish, write

$$\epsilon^{1/2} l^\epsilon(t) - l(t) = \int_0^t \frac{2}{\alpha(x^\epsilon(s))^2} dM^\epsilon - \int_0^t \frac{2}{\alpha(x^0(s))^2} dM.$$

It follows from this that

$$[\epsilon^{1/2} l^\epsilon - l]_{T\wedge\eta_r^\epsilon} \leq \|\frac{2}{\alpha}\|_{\partial D} [M^\epsilon - M]_{T\wedge\eta_r^\epsilon} + \left[\frac{2}{\alpha(x^0(\cdot))^2} - \frac{2}{\alpha(x^\epsilon(\cdot))^2}\right]_{T\wedge\eta_r^\epsilon} M^\epsilon_{T\wedge\eta_r^\epsilon}.$$

We know that the first term on the right $\to 0$ in L^2. For the second, one argues uniform integrability of $M^\epsilon_{T\wedge\eta_r^\epsilon}$ and then applies Lemma 9.1, using the continuity of $\frac{2}{\alpha(\cdot)^2}$. □

As an immediate corollary we have the following.

COROLLARY 9.1. *For any $r, T > 0, \epsilon^{1/2} l^\epsilon(T \wedge \eta_r^\epsilon)$ are uniformly integrable over all $0 < \epsilon \leq \epsilon_0$ with $e^{1/2} r < h_o$ and $y \in \partial D$.*

At last we are ready to undertake a proof of Theorem 4.1. First, by virtue of Lemma 7.2, for all $0 < \epsilon \leq \epsilon_0$,

$$\|\epsilon^{1/2} B^\epsilon[f] - \epsilon^{1/2} B^\epsilon[g]\|_{\partial D} \leq u_1(0)\|f - g\|_{\partial D}$$

and

$$\|B[f] - B[g]\|_{\partial D} \leq u_1(0)\|f - g\|_{\partial D}.$$

It therefore suffices to prove Theorem 4.1 for f in a dense subset of $C(\partial D)$. We will assume that $f \in C^1(\partial D)$. The idea is to estimate the individual terms in the following inequality, for given $r, T < \infty$.

$$|\epsilon^{1/2} B^\epsilon[f](y) - B[f](y)| \leq I + II + III + IV \tag{9.10}$$

where

$$I = \left|E_y\left[\int_{T\wedge\eta_{r^\epsilon}}^{\tau_\Gamma} f(x^\epsilon(t))\epsilon^{1/2} dl^\epsilon\right]\right|,$$

$$II = E_y\left[\int_0^{T\wedge\eta_r^\epsilon} |f(x^\epsilon(t)) - f(x^0(t))|\epsilon^{1/2} dl^\epsilon\right],$$

$$III = E_y\left[\left|\int_0^{T\wedge\eta_r^\epsilon} f(x^0(t))\epsilon^{1/2} dl^\epsilon - \int_0^{T\wedge\eta_r^\epsilon} f(x^0(t))dl\right|\right],$$

$$IV = E_y\left[\int_{T\wedge\eta_{r^\epsilon}}^{\infty} |f(x^0(t))|dl\right].$$

First consider I. Define

$$\tau^*_{\partial D} = \inf\{t \geq T \wedge \eta^\epsilon_r : x^\epsilon(t) \in \partial D\}.$$

By the strong Markov property and Lemma 7.2

$$E_y\left[\int_{T\wedge\eta^\epsilon_r}^{\tau_\Gamma} \epsilon^{1/2} dl^\epsilon\right] = E_y[E_{x^\epsilon(\tau^*_{\partial D})}[\epsilon^{1/2} l(\tau_\Gamma)],\ \tau^*_{\partial D} < \tau_\Gamma]$$

$$\leq u_1(0) P_y[\tau^*_{\partial D} < \tau_\Gamma]. \tag{9.11}$$

Now on $\{\eta^\epsilon_r \leq T\}$ the event $\{\tau^*_{\partial D} < \tau_\Gamma\}$ requires that $\varsigma^\epsilon(t)$ return from r to 0 before τ_Γ. Using Lemma 7.3, it follows that

$$P_y[\tau^*_{\partial D} < \tau_\Gamma; \eta^\epsilon_r \leq T] \leq u_2(r).$$

Thus

$$P_y[\tau^*_{\partial D} < \tau_r] \leq u_2(r) + P_y[\eta^\epsilon_r > T]$$

$$\leq u_2(r) + \frac{1}{T} u_{3,r}(0),$$

where we have appealed to Corollary 7.1. Using this in (9.11), we see that

$$I \leq \|f\|_{\partial D} u_1(0) \cdot \left[u_2(r) + \frac{1}{T} u_{3,r}(0)\right]. \tag{9.12}$$

Now consider II. If we extend f continuously from ∂D to all of $\overline{D}$, then we can write

$$II \leq E_y\left[[f(x^\epsilon(\cdot) - f(x_0(\cdot))]_T \cdot \epsilon^{1/2} l^\epsilon(T \wedge \eta^\epsilon_r)\right].$$

It follows from Lemma 9.1 that $[f(x^\epsilon) - f(x^0)]_T \to 0$ in probability, uniformly over $y \in \partial D$. This and Corollary 9.1 imply that

$$II \to 0 \text{ as } \epsilon \downarrow 0, \text{uniformly over } y \in \partial D. \tag{9.13}$$

Next consider III. Since $f \in C^1(\partial D)$, $f(x^0(t))$ is continuously differentiable. Let $\varphi(t) = d/dt\, f(x^0(t))$. We can integrate by parts to write

$$\int_0^{T\wedge\eta^\epsilon_r} f(x^0(t)) \epsilon^{1/2} dl^\epsilon = f(x^0(T \wedge \eta^\epsilon_r)) \cdot \epsilon^{1/2} l^\epsilon(T \wedge \eta^\epsilon_r) - \int_0^{T\wedge\eta^\epsilon_r} \varphi(t) \epsilon^{1/2} l^\epsilon(t) dt,$$

and likewise for the second integral in III. Therefore

$$III \leq |E_y[f(x^0(T \wedge \eta^\epsilon_r)) \cdot (\epsilon^{1/2} l^\epsilon(T \wedge \eta^\epsilon_r) - l(T \wedge \eta^\epsilon_r))]|$$

$$+ |E_y[\int_0^{T\wedge\eta^\epsilon_r} \varphi(t)[\epsilon^{1/2} l^\epsilon(t) - l(t)] dt]|$$

$$\leq (\|f\|_{\partial D} + \|\varphi\|_\infty) E_y[[\epsilon^{1/2} l^\epsilon - l]_{T\wedge\eta^\epsilon_r}].$$

Now Lemma 9.2 tells us that

$$III \to 0 \text{ as } \epsilon \downarrow 0, \text{ uniformly over } y \in \partial D. \tag{9.14}$$

Finally consider IV. Define

$$\eta^*_0 = \inf\{t \geq T \wedge \eta^\epsilon_r : \varsigma(t) = 0\}.$$

Then just as in (9.11),

$$E_y\left[\int_{T\wedge\eta^\epsilon_r}^{\infty} dl\right] = E_{0,y}[E_{0,x^0(\eta^*_0)}[l(\infty)]; \eta^*_0 < \infty] \leq u_1(0) \cdot P_{0,y}[\eta^*_0 < \infty] \tag{9.15}$$

and

$$P_{0,y}[\eta_0^* < \infty] \leq P_{0,y}[\eta_0^* < \infty; \eta_r^\epsilon \leq T] + P_{0,y}[\eta_r^\epsilon > T]. \tag{9.16}$$

From Corollary 7.1,

$$P_{0,y}[\eta_r^\epsilon > T] \leq \frac{1}{T} u_{3,r}(0). \tag{9.17}$$

Now for any $0 < \delta < r$ if $[\zeta - \zeta^\epsilon]_{T \wedge \eta_r^\epsilon} \leq \delta$ and $\eta_r^\epsilon \leq T$, then $\zeta(\eta_r^\epsilon) \geq \zeta^\epsilon(\eta_r^\epsilon) = r - \delta$, so that $\eta_{r-s} \leq \eta_r^\epsilon \leq T$. Therefore

$$P_{0,y}[\eta_0^*; \eta_r^\epsilon \leq T] \leq P_{0,y}[\eta_0^*; \eta_{r-\delta} \leq T \wedge \eta_r^\epsilon] + P_y[[\zeta - \zeta^\epsilon]_{T \wedge \eta_r^\epsilon} > \delta]. \tag{9.18}$$

The second term on the right converges to 0 as $\epsilon \downarrow 0$, uniformly over $y \in \partial D$. The first requires that $\zeta(t)$ return from $r - \delta$ to 0 in finite time. Therefore, appealing to Lemma 7.3 again

$$P_{0,y}[\eta_0^* < \infty; \eta_{r-\delta} \leq T \wedge \eta_r^\epsilon] \leq u_2(r - \delta). \tag{9.19}$$

Putting (9.15)-(9.19) together, we conclude that

$$\overline{\lim}_{\epsilon \downarrow 0} \sup_{y \in D} IV \leq \|f\|_{\partial D} \cdot u_1(0)(u_2(r - \delta) + \frac{1}{T} u_{3,r}(0)]. \tag{9.20}$$

When we put (9.12)–(9.14) and (9.20) back into (9.10) we see that, for any $r, T < \infty$ and $0 < \delta < r$,

$$\overline{\lim}_{\epsilon \downarrow 0} \|\epsilon^{1/2} B^\epsilon[f] - B[f]\|_{\partial D} \leq \|f\|_{\partial D} u_1(0) \left[u_2(r) + u_2(r - \delta) + \frac{2}{T} u_{3,r}(0) \right].$$

Finally since $u_2(r) \to 0$ as $r \to \infty$, the right side can be made arbitrarily small by choosing $\delta = 1$ and r, T sufficiently large. This proves Theorem 4.1. □

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] M. V. DAY, *Recent progress on the small parameter exit problem*, Stochastics, 20 (1987), pp. 121-150.

[3] M. V. DAY, *On the asymptotic relation between equilibrium density and exit measure in the exit problem*, Stochastics, 12 (1984), pp. 303-330.

[4] M. V. DAY, *Exponential leveling for stochastically perturbed dynamical systems*, SIAM J. Math. Anal., 13 (1982), pp. 532-540.

[5] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1984.

[6] M. I. FREIDLIN, *Functional Integration and Partial Differential Equations*, Ann. of Math. Stud. #109, Princeton University Press, Princeton, NJ, 1985.

[7] R. Z. HASMINSKII, *Stochastic Stability of Differential Equations, Sijthoff and Noordhoff*, Alphen aan den Rijn, the Netherlands, 1980.

[8] P. HSU, *Reflecting Brownian motion, boundary local time and the Neumann problem*, Ph.D. thesis, Stanford University, Stanford, CA, 1984.

[9] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North–Holland, Amsterdam, 1981.

[10] P. L. LIONS AND A. S. SZNITMAN, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1984), pp. 511-537.

[11] B. J. MATKOWSKY AND Z. SCHUSS, *Diffusion across characteristic boundaries*, SIAM J. Appl. Math., 42 (1982), pp. 822-834.

[12] B. J. MATKOWSKY AND Z. SCHUSS, *The exit problem for randomly perturbed dynamical systems*, SIAM J. Appl. Math., 33 (1977), pp. 365-382.
[13] B. J. MATKOWSKY, Z. SCHUSS, AND C. TIER, *Diffusion across characteristic boundaries with critical points*, SIAM J. Appl. Math., 43 (1983), pp. 673-695.
[14] K. SATO AND T. UENO, *Multi-dimensional diffusion and the Markov process on the boundary*, J. Math. Kyoto Univ. 4 (1965), pp. 529-605.
[15] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with boundary conditions*, Comm. Pure Appl. Math., 24 (1971), pp. 147-225.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 249-292, March 1989

A DIFFERENTIAL-DELAY EQUATION ARISING IN OPTICS AND PHYSIOLOGY*

JOHN MALLET-PARET† AND ROGER D. NUSSBAUM‡

Abstract. In recent papers the authors have studied differential-delay equations E_ε of the form $\varepsilon\dot{x}(t) = -x(t)+f(x(t-1))$. For functions like $f(x)=\mu_1+\mu_2\sin(\mu_3 x+\mu_4)$, such equations arise in optics, while for choices like $f(x)=\mu x^\nu e^{-x}$ and $f(x)=\mu x^\nu(1+x^\lambda)^{-1}$ and for $x\geqq 0$, the equation has been suggested in physiological models. Under varying hypotheses on f (labeled (I), (II), and (III) below), previous work has given theorems concerning existence and asymptotic properties as $\varepsilon\to 0^+$ of periodic solutions of E_ε which oscillate about a value α such that $f(\alpha)=\alpha$. However, verifying (I), (II), or (III) for specific examples can be difficult. This paper gives general principles that help in verifying (I), (II), or (III), and then applies these results to specific classes of functions of interest.

Key words. singularly perturbed differential-delay equation, slowly oscillating periodic solution, Schwarzian derivative

AMS(MOS) subject classifications. 26A18, 34K15, 34K25

1. Introduction.

The singularly perturbed differential-delay equation

$$\varepsilon\dot{x}(t) = -x(t)+f(x(t-1)), \tag{1.1}$$

which arises in various models in optics, biology, and physiology, has been studied by many authors. See, for example, [2], [4], [5], [7]-[14], [17]-[22], [24], [25], and the references in [20]-[22]. Recently, Mallet-Paret and Nussbaum [20]-[22] have explored the relation between (1.1) and the discrete system

$$x_n = f(x_{n-1}) \tag{1.2}$$

obtained by formally setting $\varepsilon = 0$ in (1.1). Some of the main results of [20], [21] concern the existence and asymptotic behavior of square-wavelike periodic solutions of (1.1) for small ε. However, these results require that f satisfy various hypotheses, which will be given in § 2 below and which may be nontrivial to verify. Typical nonlinearities of interest are

$$f(x) = \mu_1+\mu_2\sin(\mu_3 x+\mu_4), \tag{1.3}$$

which arise in optics, and

$$f(x) = \mu x^\nu e^{-x}, \quad x\geqq 0, \tag{1.4}$$

$$f(x) = \mu x^\nu(1+x^\lambda)^{-1}, \quad x\geqq 0, \tag{1.5}$$

which arise in biological and physiological models. See, for example, [16], where the function in (1.5) is used in (1.1) (for $\nu=0$ or $\nu=1$, $\lambda>0$ and $\mu>0$) to model blood diseases. (Note that various constants appear in the equations in [16], but that by

* Received by the editors June 30, 1987; accepted for publication March 7, 1988.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by Defense Advanced Research Projects Agency contracts 85-F-123600 and N00014-86-K-0754 and National Science Foundation grant DMS-85-07056.

‡ Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. The work of this author was supported by National Science Foundation grant DMS 85-03316.

change of variables the equations (4a) and (4b) in [16] are subsumed by our equation (1.1) with f as in (1.5).)

Unfortunately, verifying the hypotheses of § 2 even for the above simple-looking functions is not trivial and was not carried out in [21] (primarily for reasons of space). For example, one of our hypotheses involves global conditions expressed as qualitative properties of the dynamical system (1.2) and may be hard to check. It seems a significant body of theory is needed, even for the functions in (1.3)-(1.5), to determine exactly when our hypotheses are satisfied; routine calculations are insufficient. Our purpose here is to develop such a theory and then to apply it to determine parameter values for which the above nonlinearities satisfy various hypotheses. Although we have not given actual numerical ranges of parameters where our hypotheses are satisfied, we can, with a simple computer program easily obtain most of them from our results. Thus, this paper may be viewed as a companion to [21], for here we show how to apply the general results of [21] to specific systems of scientific interest.

Our interest naturally extends beyond the nonlinearities in (1.3)-(1.5); however, because so many of the basic difficulties are already apparent for these nonlinearities, we will view them as models and work out their theory in as much detail as possible. Even so, we will leave open questions for these examples.

2. Hypotheses on f and their implications. The following hypotheses were shown in [20], [21] to imply various results about the differential equation (1.1). Note that these hypotheses are arranged in increasing order of strength, and that all assume the condition $f(0)=0$. This assumption is merely a normalization; more generally the functions of interest will have a nonzero fixed point $f(x_0)=x_0$, and it will be necessary to translate this point to the origin before analyzing the function.

We say a function f is *monotone decreasing* in an interval I in case $f(x_1)\geqq f(x_2)$ whenever $x_1<x_2$ and $x_1, x_2\in I$. We say f is *strictly decreasing* in I in case $f(x_1)>f(x_2)$ for all such x_1 and x_2. We make analogous definitions of *monotone increasing* and *strictly increasing.*

We let $f^n:\mathbb{R}\to\mathbb{R}$ denote the n-fold composition of the function f with itself.

We now present four hypotheses a function f can satisfy. These were introduced in [20], [21].

(0) The function $f:\mathbb{R}\to\mathbb{R}$ is continuous, satisfies $f(0)=0$, is differentiable at $x=0$ satisfying $f'(0)<-1$, and is monotone decreasing in some neighborhood of $x=0$.

(I) The function f satisfies hypothesis (0). In addition there exist numbers $A>0$ and $B>0$ such that

$$f([-B, A])\subseteq[-B, A],$$
$$xf(x)<0 \quad \text{if } x\in[-B, A]-\{0\}.$$

(II) There exist A and B such that (I) holds. In addition there exist positive numbers $a\leqq A$ and $b\leqq B$ such that if $x_0\in[-B, A]-\{0\}$ and x_n is given by (1.2), then

$$f^n(x)=x_n\to\{-b, a\} \quad \text{as } n\to\infty.$$

(III) There exist A and B such that (I) holds. In addition f is monotone decreasing in $[-B, A]$ and (II) holds with $a=A$ and $b=B$.

Note that $f(a)=-b$ and $f(-b)=a$ must hold in hypothesis (II). If f only satisfies (I), then there still must exist a and b satisfying $f(a)=-b$ and $f(-b)=a$, respectively. However, the orbit $\{-b, a\}$ of (1.2) need not be stable and attract iterates x_n, and a and b need not be unique.

If (I) holds and f is monotone decreasing on $[-B, A]$ and if f^2 has a unique positive fixed point $a \in (0, A]$, then it is easy to show that (III) is satisfied. To see this, first observe that f^2 has a unique negative fixed point $-b \in [-B, 0)$, $-b = f(a)$. If $-b_1$ and $-b_2$ were negative fixed points of f^2, then $f(-b_1)$ and $f(-b_2)$ would be positive fixed points of f^2, so

$$a = f(-b_1) = f(-b_2),$$

and we could conclude that

$$f(a) = f^2(-b_1) = -b_1 = f^2(-b_2) = -b_2.$$

Next, note that f^2 is monotone increasing on $[-B, A]$ (because f is monotone decreasing) and that there exists $\varepsilon > 0$ such that $|f^2(x)| > x$ for $0 < |x| < \varepsilon$ (because f is monotone decreasing and $f'(0) < -1$). It follows that $f^2(x) > x$ for $0 < x < a$; otherwise, the intermediate value theorem would imply that f^2 has a positive fixed point x_1, with $0 < x_1 < a$. If $0 < y_0 < a$ and $y_n = f^{2n}(y_0)$ we conclude that

$$y_0 < f^2(y_0) = y_1 < f^2(a) = a,$$

and generally that

$$y_n < y_{n+1} < a \quad \forall n \geqq 1.$$

It follows that y_n converges to a limit y, and since $f^2(y) = y$, it must be true that $y = a$. A similar argument shows that if $-b < z_0 < 0$, then

$$\lim_{n \to \infty} f^{2n}(z_0) = -b.$$

Finally, we can deduce that if $-b \leqq x \leqq a$ and $x \neq 0$, then

$$\lim_{n \to \infty} f^n(x) = \{-b, a\}.$$

In fact we can conclude slightly more. If $A > a$, the uniqueness of the positive fixed point of f implies that $f^2(A) < A$ (we know $f^2(A) \leqq A$). Thus the intermediate value theorem implies that if $a < x \leqq A$, $f^2(x) < x$. Using this fact and the fact that f^2 is monotone increasing, we see that if $a < y_0 \leqq A$ and $y_n = f^{2n}(y_0)$, then

$$a < y_{n+1} < y_n \quad \text{for all } n.$$

As before this implies $y_n \to a$. A similar argument shows that if $-B \leqq z_0 < -b$, then

$$\lim_{n \to \infty} f^{2n}(z_0) = -b.$$

Finally, we can conclude that if $-B \leqq x \leqq A$ and $x \neq 0$, then

$$\lim_{n \to \infty} f^n(x) = \{-b, a\}.$$

If, however, f is not monotone decreasing on $[-B, A]$, then verifying (II) directly may be quite difficult, as it involves examining all iterates $x_n = f^n(x_0)$ of an arbitrary initial condition x_0. Furthermore, even if $f'(x) < 0$ for $x \in [-B, A]$, a direct proof that f^2 has exactly one fixed point in $(0, A]$ may not be easy. Fortunately, our theorems will eliminate the need for such an approach, at least in the cases of interest. Instead, checking (II) will involve only local calculations, with no need to iterate f. The main property of f that allows for such a simplification is that it possess a negative Schwarzian derivative. This property was first used in the study of interval maps by Allwright [1] and Singer [27]. If $f'(x) < 0$ for $-B < x < A$ and f has negative Schwarzian derivative on $(-B, A)$, the results of § 7 will imply f^2 has a unique fixed point in $(0, A]$.

In [21], we showed that (I), (II), and (III) each imply results about solutions of (1.1). The solutions of interest are *slowly oscillating periodic solutions*, or SOP-*solutions*. A solution $x(t)$ of (1) is called an SOP-solution if there exist quantities

$$q>1 \quad \text{and} \quad \bar{q}>q+1$$

such that

$$x(0)=x(q)=x(\bar{q})=0,$$
$$x(t)>0 \quad \text{in } (0,q),$$
$$x(t)<0 \quad \text{in } (q,\bar{q}),$$
$$x(t+\bar{q})=x(t) \quad \forall t.$$

For the functions of interest it will always be the case that $xf(x)<0$ whenever $x\neq 0$ is in the range of such a solution. In particular this will imply that the zeros of $x(t)$ are all simple.

An SOP solution $x(t)$ is called an *S-solution* if it satisfies

$$x(t+q)=-x(t) \quad \forall t,$$

in addition to the above conditions. Necessarily f is an odd function throughout the range of an S-solution. Also, $\bar{q}=2q$ for any S-solution.

The following results, which are proved in [21], describe the existence and asymptotic properties for small ε of SOP-solutions and S-solutions when (I), (II), or (III) holds.

THEOREM 2.1. *Assume f satisfies* (I). *Then there exists $\varepsilon_0>0$ such that for each positive $\varepsilon<\varepsilon_0$* (1) *possesses an* SOP-*solution satisfying*

$$x(t)\in(-B,A) \quad \forall t. \tag{2.1}$$

In addition, there exist positive numbers ε_1, γ, K_1, K_2, r_1, and r_2 such that if $x(t)$ is any SOP-*solution of* (1.1) *satisfying* (2.1), *and if $0<\varepsilon<\varepsilon_1$, then*

$$x(t)>\gamma \quad \text{for } K_2\varepsilon\leqq t\leqq q-K_2\varepsilon,$$
$$x(t)<-\gamma \quad \text{for } q+K_2\varepsilon\leqq t\leqq\bar{q}-K_2\varepsilon,$$
$$|\dot{x}(t)|\geqq K_1/\varepsilon \quad \text{whenever } |x(t)|\leqq\gamma,$$
$$1+\varepsilon r_1\leqq q\leqq 1+\varepsilon r_2,$$
$$1+\varepsilon r_1\leqq\bar{q}-q\leqq 1+\varepsilon r_2.$$

THEOREM 2.2. *Assume f satisfies* (II). *Then given $\delta>0$ there exist $\varepsilon_2>0$ and $K_2>0$ such that if $x(t)$ is any* SOP-*solution of* (1) *satisfying* (2.1), *and if $0<\varepsilon<\varepsilon_2$, then*

$$|x(t)-\mathrm{sqw}\,(t)|\leqq\delta \quad \text{in } [\varepsilon K_2, q-\varepsilon K_2]\cup[q+\varepsilon K_2,\bar{q}-\varepsilon K_2]$$

where sqw (t) *is the two-periodic square-wavefunction defined by*

$$\mathrm{sqw}(t)=\begin{cases} a & \text{in } [0,1),\\ -b & \text{in } [1,2),\end{cases}$$
$$\mathrm{sqw}\,(t+2)=\mathrm{sqw}\,(t) \quad \forall t.$$

THEOREM 2.3. *Assume f satisfies* (III). *Let $x(t)$ be any* SOP-*solution of* (1) *satisfying* (2.1) *for some $\varepsilon>0$ (with $a=A$ and $b=B$), and let $p\in(0,q)$ and $\bar{p}\in(q,\bar{q})$ be such that*

$$x(p)=\max x(t) \quad \text{and} \quad x(\bar{p})=\min x(t).$$

Then $x(t)$ is monotone increasing in $(0, p)$, monotone decreasing in $(p, \bar{p})$, and monotone increasing in $(\bar{p}, \bar{q})$.

THEOREM 2.4. *Assume f satisfies* (I) *and that in addition $A = B$ and $f(-x) = -f(x)$ for all $x \in [-A, A]$. Then there exists $\varepsilon_0 > 0$ such that for each positive $\varepsilon < \varepsilon_0$,* (2.1) *possesses an S-solution satisfying*

$$x(t) \in (-A, A) \quad \forall t.$$

In the case of Theorem 2.2 we easily see that

$$x(t) \to \operatorname{sqw}(t) \quad \text{as } \varepsilon \to 0$$

uniformly on compact subsets of $\mathbb{R} - \mathbb{Z}$, for SOP-solutions $x(t)$. Also, when f is odd the S-solutions obtained in Theorem 2.4 are of course SOP-solutions, and hence satisfy the conclusions of Theorems 2.1, 2.2, and 2.3 when the appropriate hypotheses hold.

3. Some specific functions f. We consider $f_k : \mathbb{R} \to \mathbb{R}$, for $1 \leqq k \leqq 5$, defined as follows:

$$f_1(x) = \mu - x^2,$$

$$f_2(x) = x^3 - \mu x,$$

$$f_3(x) = -\mu[\sin(x + \theta) - \sin\theta],$$

$$f_4(x) = \mu x^\nu e^{-x}, \qquad x \geqq 0,$$

$$f_5(x) = \frac{\mu x^\nu}{x^\lambda + 1}, \qquad x \geqq 0.$$

The values of f_4 and f_5 for $x < 0$ are immaterial, so for definiteness we set

$$f_k(x) = f_k(0) \quad \text{if } x < 0 \text{ and } k = 4 \text{ or } 5,$$

always assuming $\nu \geqq 0$ and $\lambda \geqq 0$. The functions f_1 and f_2 give model problems with the simplest possible nonlinearities; in particular f_1 is the much-studied quadratic map of the interval [6], [15]. The function f_2 is an odd function, so by Theorem 2.4 there is the possibility of obtaining S-solutions of (1.1). The function f_3 seems at first to be a special case of the general trigonometric nonlinearity (1.3) arising in optical models; we will show, however, that f_3 can always be obtained from (1.3) by means of a linear transformation of the differential equation (1.1). The function f_4 occurs in biological and physiological models as noted earlier, as does f_5 when $\nu = 1$ or $\nu = 0$.

Our object is to determine ranges of the parameters μ, θ, ν, and λ for which the hypotheses (0), (I), (II), and (III) hold for a suitable translate of each f_k. By "suitable translate" we mean that a transformation taking a fixed point x_0 of f_k to the origin must generally be made before verifying the hypothesis in question. Indeed, for f_4 and f_5 it is not the fixed point $x = 0$ that is of interest, but rather some nontrivial fixed point $x_0 > 0$ about which we do our analysis. If $f : \mathbb{R} \to \mathbb{R}$ possesses a fixed point x_0, then letting $y = x - x_0$ in (1) yields

$$\varepsilon \dot{y}(t) = -y(t) + g(y(t-1)) \tag{3.1}$$

where

$$g(y) = f(y + x_0) - f(x_0) \tag{3.2}$$

satisfies $g(0) = 0$. When we say *a hypothesis* (such as (0), (I), (II), or (III)) *holds for a function f at a fixed point x_0*, we mean that the hypothesis holds for the transformed function g as stated.

We complete this section by showing how the function f in (1.3) can be reduced to the normal form f_3. In fact, we will show that the parameters μ and θ can always be chosen to satisfy

$$\mu \geqq 0 \quad \text{and} \quad 0 \leqq \theta \leqq \pi. \tag{3.3}$$

First note that the function f in (1.3) is bounded and so must have at least one fixed point, and possibly more than one. Let x_0 denote such a point. Then the function g in (3.2) has the same form as in (1.3) but possibly with a different value of μ_4; we continue to denote the new value by μ_4. The fact that $g(0)=0$ implies from the form (1.3) that $\mu_1 = -\mu_2 \sin \mu_4$, and so

$$g(y) = \mu_2[\sin(\mu_3 y + \mu_4) - \sin \mu_4].$$

Now assuming $\mu_2 \neq 0$ and $\mu_3 \neq 0$ (otherwise g is identically zero), we set

$$z = \pm \mu_3 y \tag{3.4}$$

with the sign $\pm$ to be determined later. The differential equation (3.1) now becomes

$$\varepsilon \dot{z}(t) = -z(t) + h(z(t-1))$$

where

$$h(z) = \mu_2 \mu_3[\sin(z \pm \mu_4) - \sin(\pm \mu_4)].$$

Upon setting

$$\mu = |\mu_2 \mu_3|,$$

$$\theta = \begin{cases} \pm \mu_4 (\bmod 2\pi) & \text{if } \mu_2 \mu_3 < 0, \\ \pm \mu_4 + \pi (\bmod 2\pi) & \text{if } \mu_2 \mu_3 > 0, \end{cases}$$

we see that the function h has precisely the form of f_3. In addition, an appropriate choice of sign in (3.4) ensures that μ and θ satisfy (3.3).

In our subsequent analysis we will usually assume that the function $f(x)$ in (1.3) has been written in the normal form:

$$f_3(x) = -\mu[\sin(x+\theta) - \sin \theta] \tag{3.5}$$

with $\mu > 0$ and $0 \leqq \theta \leqq \pi$. However, the reader should remember that writing the function in normal form conceals certain difficulties. First, as previously noted, the function f in (1.3) may have several fixed points. For each such fixed point of f, different parameters μ and θ in the normal form f_3 will, in general, be obtained. Second, we usually want to know for what ranges of the *original* parameters μ_1, μ_2, μ_3, and μ_4 in (1.3) does the function $f(x)$ satisfy hypotheses (0), (I), (II), or (III). The parameters in the normal form are written in terms of a fixed point of f in (1.3), and this fixed point is typically not explicitly known. Thus transferring information about the normal form back to the original function may present some nontrivial calculus problems.

4. The local condition (0). Here we discuss the existence of a fixed point of f_k at which condition (0) holds; clearly this is the case at a fixed point x_0 if and only if $f_k'(x_0) < -1$. For each function f_k, with parameters θ, ν, and λ in appropriate ranges, we will show that a critical value μ_0 of the parameter μ exists such that (0) holds at an appropriate fixed point x_0 if and only if $\mu > \mu_0$.

We begin with the model functions f_1 and f_2. If $\mu > -1/4$ then f_1 has two fixed points; the larger one,

$$x_0 = \frac{-1+\sqrt{4\mu+1}}{2},$$

interests us here. We see that $f_1'(x_0) = 1-\sqrt{4\mu+1}$, and a short calculation reveals that (0) holds there if and only if $\mu > \mu_0 = \frac{3}{4}$. For the nonlinearity f_2, with the fixed point $x_0 = 0$, we have $f_2'(0) = -\mu$; thus (0) holds there if and only if $\mu > \mu_0 = 1$.

At the fixed point $x_0 = 0$ of f_3, we have $f_3'(0) = -\mu \cos\theta$, so a necessary and sufficient condition for (0) to hold here is that $\mu \cos\theta > 1$. In particular this condition and the restrictions (3.3) imply that $\mu > 0$ and $0 \leqq \theta < \pi/2$. Thus we obtain $\mu_0 = 1/\cos\theta$.

Before discussing the functions f_4 and f_5 it is convenient to prove a simple theorem.

THEOREM 4.1. *Let $\bar{f}:[0,\infty)\to[0,\infty)$ be a continuous function that is C^1 on $(0,\infty)$. Assume there exists $\theta \geqq 0$ such that $\bar{f}'(x) > 0$ for $0 < x < \theta$ and $\bar{f}'(x) < 0$ for $x > \theta$, and there exists $s_0 > 0$ such that $(d/dx)(x\bar{f}(x))$ is positive for $0 < x < s_0$ and negative for $x > s_0$. Then for $\mu \geqq \theta(\bar{f}(\theta))^{-1}$, the equation $\mu\bar{f}(x) = x$ has a unique solution $x = x_0(\mu)$ such that $x_0(\mu) \geqq \theta$ and $\mu\bar{f}$ satisfies (0) at $x_0(\mu)$ if and only if $\mu > s_0(\bar{f}(s_0))^{-1}$.*

Proof. The existence and uniqueness of $x_0(\mu)$ is trivial. Since $s(\bar{f}(s))^{-1}$ is strictly increasing for $s > \theta$, we can define $\mu(s) = s(\bar{f}(s))^{-1}$ and parameterize by $s \geqq \theta$, so $\mu(s)\bar{f}(s)$ has fixed point $s \geqq \theta$. Thus the set of μ such that $\mu f'(x_0(\mu)) < -1$ is the same as $\{\mu(s): \mu(s)\bar{f}'(s) < -1\}$. A calculation shows that $\mu(s)\bar{f}'(s) < -1$ if and only if $(d/ds)(s\bar{f}(s)) < 0$, i.e., if and only if $s > s_0$. □

For the function $\bar{f}_4(x) = x^\nu e^{-x}$ we easily compute that the conditions of Theorem 4.1 are satisfied for $\nu \geqq 0$ and that $s_0 = \nu+1$ and $\mu\bar{f}_4(x)$ satisfies (0) at x_0 if and only if $\mu > (\nu+1)(\bar{f}_4(\nu+1))^{-1}$. For the function $\bar{f}_5(x) = x^\nu(1+x^\lambda)^{-1}$, we easily compute that the hypotheses of Theorem 4.1 are satisfied if $\nu \geqq 0$ and $\lambda > \nu+1$ and that $s_0^\lambda = (\nu+1)(\lambda-\nu-1)^{-1}$. Thus $\mu\bar{f}_5(x)$ satisfies condition (0) at x_0 if and only if $\mu > s_0(\bar{f}_5(s_0))^{-1}$, where $s_0^\lambda = (\nu+1)(\lambda-\nu-1)^{-1}$.

Table 1 summarizes the previous results by giving the range of parameters for which f_k satisfies (0) at a fixed point x_0. Note again that in the case of f_3 only the point $x_0 = 0$ is considered, even though there may be other fixed points at which (0) holds.

TABLE 1

Fixed points x_0 of f_k and critical parameter values μ_0. We have (0) holding at x_0 if and only if $\mu > \mu_0$, provided the parameters θ, ν, and λ satisfy the given restrictions.

k	x_0	μ_0	Restrictions
1	$\frac{-1+\sqrt{4\mu+1}}{2}$	$\frac{3}{4}$	
2	0	1	
3	0	$\frac{1}{\cos\theta}$	$0 \leqq \theta < \frac{\pi}{2}$
4	unique fixed point $x_0 > \nu$	$(\nu+1)^{1-\nu}e^{\nu+1}$	$\nu \geqq 0$
5	unique fixed point $x_0 > \left(\frac{\nu}{\lambda-\nu}\right)^{1/\lambda}$	$\log\mu_0 = \log\lambda - \left(\frac{\nu-1}{\lambda}\right)\log(\nu+1) - \left(\frac{\lambda-\nu+1}{\lambda}\right)\log(\lambda-\nu-1)$	$\nu \geqq 0, \quad \lambda > \nu+1$

When $\nu = 1$ in the function $f_5(x)$, we can explicitly compute $x_0 = (\mu - 1)^{1/\lambda}$, and the conditions on the parameters become $\lambda < 2$ and $\mu > \lambda(\lambda - 2)^{-1}$. More generally, it is of interest to locate the fixed points of f_4 and f_5 more precisely. The following result gives the asymptotic behavior of $x_0(\mu)$ for large μ; we omit the proof because we will not actually use the result here and because the proof involves only standard arguments from asymptotic analysis. Note that we use the standard "big O" and "little o" notation: If $h(\mu)$ and $g(\mu)$ are complex-valued functions defined for large positive μ and if $g(\mu)$ is nonzero for large μ, then we write

$$h(\mu) = O(g(\mu)) \quad \text{if and only if } \limsup_{\mu \to +\infty} \frac{|h(\mu)|}{|g(\mu)|} < \infty,$$

$$h(\mu) = o(g(\mu)) \quad \text{if and only if } \limsup_{\mu \to +\infty} \frac{|h(\mu)|}{|g(\mu)|} = 0.$$

THEOREM 4.2. *For $\nu \geqq 0$ and $\lambda > \nu + 1$ define numbers $\theta_4 = \nu$ and $\theta_5^{\lambda} = \nu(\lambda - \nu)^{-1}$. The functions $f_4(x) = \mu \bar{f}_4(x)$ and $f_5(x) = \mu \bar{f}_5(x)$ have (for sufficiently large μ) a unique fixed point $x_0(\mu)$ such that $x_0(\mu) > \theta_j$ and $x_0(\mu)$ satisfies*

$$x_0(\mu) = \log(\mu) - (1 - \nu)\log(\log \mu) + O\left(\frac{\log(\log \mu)}{\log \mu}\right) \quad \text{for } f_4,$$

$$x_0(\mu) = \mu^{(1/(\lambda+1-\nu))} - \left(\frac{1}{\lambda + 1 - \nu}\right)\mu^{((1-\lambda)/(\lambda+1-\nu))} + O(\mu^{((1-2\lambda)/(\lambda+1-\nu))}) \quad \text{for } f_5,$$

where log *denotes natural logarithm.*

5. General results on piecewise monotone functions. We wish to determine when a hypothesis (I), (II), or (III) holds for f_k at a point x_0 given in Table 1. As noted earlier these three conditions are global, and verifying them for specific functions may be difficult. To aid us in this task we will first obtain some general criteria for these hypotheses to hold; we will then apply these criteria to the functions f_k of interest.

To begin, we introduce condition (PM) (piecewise monotone) on a function f; observe that each f_k satisfies (PM) at the fixed point x_0 and parameter ranges of Table 1:

(PM) *The function $f: \mathbb{R} \to \mathbb{R}$ satisfies hypothesis* (0). *In addition, there exist (possibly infinite) quantities $0 < \xi \leqq \alpha \leqq \infty$ and $0 < \eta \leqq \beta \leqq \infty$ such that*

(i) *$f(-\beta) = 0$ if $\beta < \infty$;*

(ii) *f is monotone increasing and strictly positive in $(-\beta, -\eta)$ if $\eta < \infty$;*

(iii) *f is monotone decreasing in $(-\eta, \xi)$ but not in any larger open interval;*

(iv) *f is monotone increasing and strictly negative in (ξ, α) if $\xi < \infty$;*

(v) *$f(\alpha) = 0$ if $\alpha < \infty$.*

Furthermore, if $\xi = \eta = \infty$ (so f is monotone decreasing in all of $\mathbb{R}$), then $|f^2(x)| < |x|$ for some x.

Figure 1 depicts a function satisfying (PM). (Note that $f(\alpha) = 0$ is *not* required for this function as $\alpha = \infty$. That is, $\lim_{x \to \infty} f(x)$ may either be zero or strictly negative.) For any function satisfying (PM) we have $xf(x) < 0$ if $|x| \neq 0$ is sufficiently small, since $f(0) = 0$ and $f'(0) < -1$ by (0). A first question we consider for such a function is when it also satisfies (I).

Suppose both (PM) and (I) hold for f. Then the quantities A and B in (I) clearly satisfy

$$A < \alpha \quad \text{and} \quad B < \beta. \tag{5.1}$$

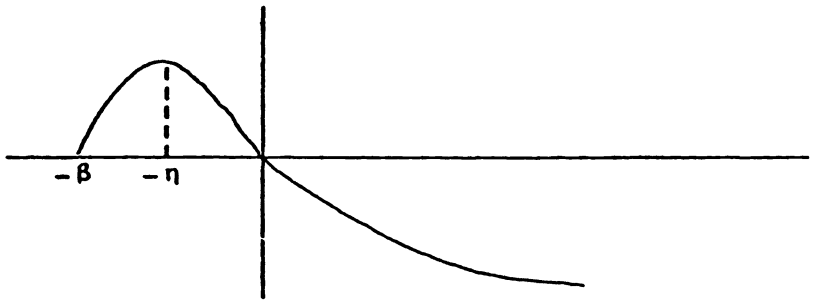


FIG. 1

On the other hand, if f satisfies (PM) and A and B are positive numbers satisfying (5.1), then f also satisfies (I) if and only if

$$f([-B, A]) \subseteq [-B, A]. \tag{5.2}$$

Finally, if both (5.1) and (5.2) hold for a function satisfying (PM), then there exists a minimal interval $[-B_*, A_*] \subseteq (-\beta, \alpha)$, with both A_* and B_* strictly positive, for which (5.2) is an equality:

$$f([-B_*, A_*]) = [-B_*, A_*].$$

This is so because $f'(0) < -1$.

Now assume f satisfies (PM) and define a continuous monotone decreasing function $f_*: R \to R$ by

$$f_*(x) = \begin{cases} f(-\eta) & \text{in } (-\infty, -\eta] \quad \text{if } \eta < \infty, \\ f(x) & \text{in } (-\eta, \xi), \\ f(\xi) & \text{in } [\xi, \infty) \quad \text{if } \xi < \infty. \end{cases} \tag{5.3}$$

If A and B are positive numbers satisfying (5.1), then clearly

$$f([-B, A]) = [f_*(A), f_*(-B)], \tag{5.4}$$

so that the equality in (5.2) holds at $A = A_*$ and $B = B_*$ if and only if

$$f_*(A_*) = -B_* \quad \text{and} \quad f_*(-B_*) = A_*. \tag{5.5}$$

Thus, (I) holds if and only if there exist two distinct points in the interval $(-\beta, \alpha)$ that are mapped into one another by f_*. From this basic fact we conclude the following result.

PROPOSITION 5.1. *Assume f satisfies* (PM), *define the function f_* by* (5.3), *and define quantities*

$$\begin{aligned} A_* &= \inf \{A > 0 \mid f_*^2(A) = A\}, \\ B_* &= \inf \{B > 0 \mid f_*^2(-B) = -B\}. \end{aligned} \tag{5.6}$$

Then A_* and B_* are well-defined positive numbers satisfying (5.5). *Hypothesis* (I) *holds for f if and only if*

$$A_* < \alpha \quad \text{and} \quad B_* < \beta$$

And in such a case we can take $A = A_*$ and $B = B_*$ in the statement of (I). *A sufficient condition for* (I) *to hold is that both*

$$f_*^2(\alpha) < \alpha \quad \text{if} \quad \alpha < \infty, \qquad f_*^2(-\beta) > -\beta \quad \text{if } \beta < \infty. \tag{5.7}$$

Another sufficient condition for (I) *to hold is that*

$$f_*(\alpha) > -\beta, \quad f_*^2(\alpha) < \alpha, \quad \alpha < \infty, \tag{5.8}$$

while a third sufficient condition for (I) *to hold is that*

$$f_*(-\beta) < \alpha, \quad f_*^2(-\beta) > -\beta, \quad \beta < \infty. \tag{5.9}$$

Proof. The existence and positivity of A_* and B_* follow from the fact that $|f_*^2(x)| > |x|$ for small $|x| \neq 0$ (since $(f_*^2)'(0) > 1$) and from the fact that $|f_*^2(x)| < |x|$ for some x. The latter inequality holds because f_*^2 is a bounded function if either ξ or η is finite; if $\xi = \eta = \infty$, the inequality is assumed in the definition of (PM). The first part of the proposition now follows easily from the monotonicity of f_* and the discussion above.

The assumptions on f imply that f_*^2 always has a positive fixed point and a negative fixed point. If $\alpha = \beta = \infty$, f satisfies (I) by what we have already proved. In all other cases it suffices to prove that f_*^2 has a fixed point in $(0, \alpha)$ and a fixed point in $(-\beta, 0)$. The reader can easily verify that (5.7), (5.8), or (5.9) are all sufficient to ensure this. For example, if $\alpha < \infty$ and (5.8) is satisfied, f_*^2 has a fixed point x_0 in $(0, \alpha)$ because $f_*^2(\alpha) < \alpha$ and $(f_*^2)'(0) > 1$, and then $f_*(x_0)$ is a fixed point of f_*^2 in $(-\beta, 0)$ (because f_* is monotone and $f_*(\alpha) > -\beta$). □

The following related result tells when the monotonicity condition (III) holds for a function that satisfies (PM).

PROPOSITION 5.2. *Assume f satisfies* (PM), *and let f_*, A_*, and B_* be as in Proposition* 5.1. *Then* (III) *holds for f if and only if*

$$A_* \leqq \xi \quad \text{and} \quad B_* \leqq \eta. \tag{5.10}$$

As before, we have $A = A_*$ and $B = B_*$ in the statement of (III). *A sufficient condition for* (5.10) *to hold is that both*

$$f_*^2(\xi) \leqq \xi \quad \text{if } \xi < \infty, \qquad f_*^2(-\eta) \geqq -\eta \quad \text{if } \eta < \infty \tag{5.11}$$

should hold.

Another sufficient condition for (5.10) *to be satisfied is that*

$$f_*(\xi) \geqq -\eta, \quad f_*^2(\xi) \leqq \xi, \quad \xi < \infty \tag{5.12}$$

while a third sufficient condition for (5.10) *to hold is that*

$$f_*(-\eta) \leqq \xi, \quad f_*^2(-\eta) \geqq -\eta, \quad \eta < \infty. \tag{5.13}$$

Proof. This follows the proof of Proposition 5.1 once we recall f is monotone in $(-\eta, \xi)$, but in no larger open interval. □

Suppose the function f_*^2 has exactly one fixed point in $(0, \infty)$. Then an easy argument implies that f_*^2 has a unique fixed point in $(-\infty, 0)$. If $A_* < \alpha$ and $B_* < \infty$ (A_* and B_* as in (5.6)), we must have $f_*^2(\alpha) < \alpha$ (if $\alpha < \infty$) and $f_*^2(-\beta) > -\beta$ (if $\beta < \infty$): otherwise the intermediate value theorem would imply that f_*^2 has a fixed point in $[\alpha, \infty)$ or $(-\infty, -\beta]$, contradicting uniqueness. Thus if f_*^2 has a unique positive fixed point, the sufficient condition in (5.7) that f satisfy (I) is also necessary. Furthermore, a little additional thought shows that (5.7) is satisfied if and only if (5.8) is satisfied, or (5.9) is satisfied, or $\alpha = \beta = \infty$ (assuming f_*^2 has a unique positive fixed point).

Similarly, if f is as in Proposition 5.2 and f_*^2 has a unique positive fixed point, then f satisfies (III) if and only if f satisfies (5.11). Also f satisfies (III) if and only if f satisfies (5.12), or (5.13), or $\xi = \eta = \infty$.

Inequalities (5.7) and (5.11) are, in general, easier to verify than (5.6) and (5.10), so it is useful to have theorems that ensure f_*^2 has a unique positive fixed point. Furthermore, we have already seen in the discussion in § 2 the importance of knowing that $f^2:[-B, A]\to[-B, A]$ (A and B as in § 2) has a unique positive fixed point. For example, if f is monotone decreasing and $f([-B, A])\subset[-B, A]$, we saw in § 2 that f satisfies (III) on $[-B, A]$ if f^2 has a unique positive fixed point in $[-B, A]$.

If the function f_*^2 were convex downward in (τ, ∞) and convex upward in $(-\infty, \tau)$ for some real τ, then f_*^2 would have a unique fixed point in $(0, \infty)$. This type of convexity assumption is clumsy to deal with, but a related concept, that of negative Schwarzian derivative, is readily verifiable for many functions of interest and can be used to prove a variety of results, including the uniqueness of the positive fixed point of f_*^2. Remarkably, each of the five functions f_k has a negative Schwarzian derivative for most of the parameter values of interest, so it will be natural for us to make the assumption of negative Schwarzian derivative in most of our subsequent theorems.

6. The Schwarzian derivative. The Schwarzian derivative Sf of a function $f: I\to R$ in an interval I is defined to be the function

$$Sf(x)=\frac{f'''(x)}{f'(x)}-\frac{3}{2}\left(\frac{f''(x)}{f'(x)}\right)^2$$

at those points $x\in I$ where f is three times differentiable and $f'(x)\neq 0$. At all other points of I we consider Sf to be undefined. The Schwarzian derivative originated in the theory of conformal mappings and was first used in the study of interval maps by Allwright [1] and Singer [27].

In this section we present several basic properties of Schwarzian derivatives and of functions whose Schwarzian derivative is negative. An important sufficient condition for the Schwarzian derivative of a function to be negative is given in Proposition 6.2.

Our first proposition collects some well-known results about the Schwarzian derivative (see [6], [27]). Statement (v) in Proposition 6.1, although elementary, is quite useful and does not appear to have been explicitly stated in the literature.

PROPOSITION 6.1. *Let $f: I\to\mathbb{R}$ and $y: J\to\mathbb{R}$ be functions defined on intervals I and J. Then*

(i) $S(g\circ f)(x)=Sf(x)+[f'(x)]^2 Sg(f(x))$ *and*

(ii) $S(g\circ f)(x)<0$ *if* $Sf(x)<0$ *and* $Sg(f(x))<0$

hold whenever Sf is defined at $x\in I$ and Sg is defined at $f(x)\in J$. Also, if m is the Möbius transformation

$$m(x)=\frac{c_1x+c_2}{c_3x+c_4},\qquad c_1c_4-c_2c_3\neq 0,$$

then

(iii) $Sm(x)=0$ *where defined, and hence*

$$S(m\circ f)(x)=Sf(x)$$

if Sf is defined at $x\in I$ and $c_3f(x)+c_4\neq 0$. Finally,

(iv) $(d^2/dx^2)(|f'(x)|^{-1/2})>0$ *if and only if* $Sf(x)<0$, *and*

(v) *if* $(d^2/dx^2)\log|f'(x)|<0$ *then* $Sf(x)<0$

hold whenever Sf is defined.

Proof. These are straightforward but tedious calculations, which we omit. We do note that (ii) and (iii) follow easily from (i). Also,

$$h''(x) > 0 \quad \text{implies } \frac{d^2}{dx^2}(e^{h(x)}) > 0;$$

so with $h(x) = -\frac{1}{2}\log|f'(x)|$ we obtain (v) from (iv). □

COROLLARY 6.1 [27, Prop. 2.4]. *Let $f: I \to \mathbb{R}$ be three times differentiable in an open interval I, and assume*

$$f'(x) \neq 0 \quad \textit{and} \quad Sf(x) < 0 \quad \textit{for all } x \in I.$$

Then the function $|f'(x)|$ does not attain a minimum in I. That is, there does not exist $w \in I$ such that $|f'(w)| \leqq |f'(x)|$ for all $x \in I$.

Proof. This follows immediately from (iv) of Proposition 6.1. □

LEMMA 6.1. *Let $f: I \to I$ be three times differentiable in an interval I (not necessarily open), with range in I, and assume that $Sf(x) < 0$ at each $x \in I$ for which $f'(x) \neq 0$. Assume that for some $w \in I$ (possibly an endpoint) we have*

$$f(w) = w \quad \textit{and} \quad |f'(w)| \leqq 1.$$

In addition, assume

$$f''(w) = 0 \quad \textit{if } f'(w) = 1$$

at this fixed point w. Then there exists a relatively open neighborhood $U \subseteq I$, with $w \in U$, such that

$$f(U) \subseteq U,$$

$$f^n(x) \to w \quad \textit{as } n \to \infty, \quad \textit{for each } x \in U.$$

Proof. In the case of a strict inequality $|f'(w)| < 1$ it is an elementary exercise to show that the set

$$U = (w - \delta, w + \delta) \cap I$$

satisfies the conclusions of the lemma if $\delta > 0$ is small enough. The same set also works if $f'(w) = 1$: we have $f''(w) = 0$ (by assumption) and $f'''(w) < 0$ (since $Sf(w) < 0$). It follows that if $g(x) = f(x) - x$, $g^{(j)}(w) = 0$ for $0 \leqq j \leqq 2$ and $g^{(3)}(w) < 0$. Thus Taylor's theorem implies that there exists $\delta > 0$ such that $f(x) < x$ for $x \in (w, w + \delta) \cap I$ and $f(x) > x$ for $x \in (w - \delta, w) \cap I$. Since $f'(w) = 1$ we also have $f(x) \geqq w$ for $x \in (w, w + \delta) \cap I$ and $f(x) \leqq w$ for $x \in (w - \delta, w) \cap I$. It follows that if $x \in (w - \delta, w + \delta) \cap I$, $x_n = f^n(x)$ is a monotonic sequence bounded above by w (if $x \leqq w$) or below by w (if $w \leqq x$). Thus the sequence (x_n) converges to ξ and necessarily $f(\xi) = \xi$. By construction w is the only fixed point of f in $(w - \delta, w + \delta) \cap I$, so $\xi = w$ and x_n converges to ξ.

In the remaining case, when $f'(w) = -1$, we have $f^2(w) = w$ and $(f^2)'(w) = 1$, and a simple calculation (see [27, p. 261]) gives us that $(f^2)''(w) = 0$. By (ii) of Proposition 6.1 the Schwarzian derivative of f^2 satisfies $Sf^2(x) < 0$ whenever $(f^2)'(x) \neq 0$. Thus f^2 satisfies the conditions on f already considered in the preceding paragraph, so there exists $U_0 \subseteq I$ satisfying the conclusions of the lemma for the function f^2 instead of f. From this we see that the set $U = U_0 \cap f(U_0)$ satisfies the conclusions of the lemma for the function f. □

The next result gives an easily verified condition for the Schwarzian derivative of a function to be negative. Recall first that the order ω of an entire function $f: \mathbb{C} \to \mathbb{C}$ is the infimum of all numbers $\kappa > 0$ such that $|f(z)|\, e^{-|z|^\kappa}$ is bounded on $\mathbb{C}$. (If no such

κ exists then f is said to have infinite order.) Nontrivial entire functions of finite order possess a product representation

$$f(z) = e^{\Omega(z)} z^k \prod_n \left(1 - \frac{z}{z_n}\right) E_M\left(\frac{z}{z_n}\right)$$

with at most countably many factors, where Ω is a polynomial of degree at most $[\omega]$ (the greatest integer less than or equal to the order), $k \geqq 0$ is the multiplicity of $z = 0$ as a root of f, the numbers z_n are the other roots of f listed according to multiplicity, E_M is the function

$$E_M(z) = \exp\left(\sum_{n=1}^{M} \frac{z^n}{n}\right)$$

with $E_0(z) = 1$, and $M \geqq 0$ is an integer satisfying $M \leqq \omega \leqq M+1$. In addition, it is the case that

$$\sum_n \frac{1}{|z_n|^{M+1}} < \infty, \tag{6.1}$$

and the infinite product converges uniformly on compact subsets of $\mathbb{C}$. We also recall that the order of the derivative f' equals the order ω of f.

PROPOSITION 6.2. *Let f be an entire function of order $\omega < 2$ such that $f(x) \in \mathbb{R}$ whenever $x \in \mathbb{R}$, and such that all zeros of the derivative f' are real. Then either*

$$Sf(x) < 0 \quad \textit{whenever } f'(x) \neq 0 \textit{ and } x \in \mathbb{R},$$

or f is a linear function $f(x) = c_0 x + c_1$.

Proof. We note that $M = 0$ or 1 in the infinite product representation for the derivative f'. Denoting the zeros of f' by $x_n \in \mathbb{R}$ we have for this function either

$$f'(x) = e^{\Omega(x)} x^k \prod_n \left(1 - \frac{x}{x_n}\right), \quad \text{or} \tag{6.2}$$

$$f'(x) = e^{\Omega(x)} x^k \prod_n \left(1 - \frac{x}{x_n}\right) e^{x/x_n}, \tag{6.3}$$

where we restrict our attention to real values x of the argument. In either case (6.2) or (6.3) we have $\Omega''(x) = 0$ for all x, and

$$\sum_n \frac{1}{|x_n|^2} < \infty. \tag{6.4}$$

Now assume that f' does possess a root; otherwise $f'(x) = e^{\Omega(x)}$ and the result is easily checked. In the first case, (6.2), we have

$$\log|f'(x)| = \Omega(x) + k \log|x| + \sum_n \log\left|1 - \frac{x}{x_n}\right|,$$

so term-by-term differentiation (justified by (6.1)*) gives*

$$\frac{d^2}{dx^2} \log|f'(x)| = -\frac{k}{x^2} - \sum_n \frac{1}{(x - x_n)^2} < 0 \tag{6.5}$$

for $x \neq x_n$ and $x \neq 0$ if $k > 0$. In the second case, (6.3), the same formula, (6.5), holds. In either case the result $Sf(x) < 0$ follows from (v) of Proposition 6.1. □

7. Verifying (I), (II), and (III). The following hypotheses are strengthened versions of (PM) involving a negative Schwarzian derivative condition. Under these conditions, verifying (I), (II), and (III) reduces to essentially local calculations.

(NS_1) *The function* $f:\mathbb{R}\to\mathbb{R}$ *satisfies hypothesis* (PM). *In addition* f *is three times differentiable in* $(-\eta, \xi)$ *and satisfies* $f'(x)<0$ *and* $Sf(x)<0$ *in* $(-\eta, \xi)$.

(NS_2) *The function* $f:\mathbb{R}\to\mathbb{R}$ *satisfies* (PM). *In addition,* f *is three times differentiable in* $(-\beta, \alpha)$ *and satisfies* $f'(x)\neq 0$ *and* $Sf(x)<0$ *if* $x\in(-\beta, \alpha)$ *and* $x\neq \xi, -\eta$.

Under hypothesis (NS_1), Theorem 7.1 gives necessary and sufficient condition for (I) and (III) to hold, thereby extending Propositions 5.1 and 5.2. Under hypothesis (NS_2), Theorem 7.3 gives an easily verified necessary and sufficient condition for (II) to hold provided (I) also holds. Theorem 7.2 will be useful in verifying (I), (II), and (III) for specific functions.

THEOREM 7.1. *Assume* f *satisfies* (NS_1) *and define* f_* *by* (5.3) *as before. Then* f *satisfies* (I) *if and only if*

$$f_*^2(\alpha)<\alpha \quad \text{if } \alpha<\infty, \qquad f_*^2(\beta)>-\beta \quad \text{if } \beta>\infty. \tag{7.1}$$

Also, f *satisfies* (I) *if and only if at least one of the following three conditions holds*:

(1) $\alpha=\beta=\infty$; *or* (2) $\alpha<\infty$ *and* $f_*(\alpha)>-\beta$ *and* $f_*^2(\alpha)<\alpha$; *or* (3) $\beta<\infty$ *and* $f_*(-\beta)<\alpha$ *and* $f_*^2(-\beta)>-\beta$.

The function f *satisfies* (III) *if and only if*

$$f_*^2(\xi)\leqq\xi \quad \text{if } \xi<\infty, \qquad f_*^2(-\eta)\geqq-\eta \quad \text{if } \eta<\infty. \tag{7.2}$$

Also, f *satisfies* (III) *if and only if at least one of the following three conditions is satisfied*:

(1) $\xi=\eta=\infty$; *or* (2) $\xi<\infty$ *and* $f_*(\xi)\geqq-\eta$ *and* $f_*^2(\xi)\leqq\xi$; or (3) $\eta<\infty$ *and* $f_*(-\eta)\leqq\xi$ *and* $f_*^2(-\eta)\geqq-\eta$.

Recall that we have

$$f_*^2(\alpha)=f_*^2(\xi) \quad \text{and} \quad f_*^2(-\beta)=f_*^2(-\eta) \tag{7.3}$$

for the quantities in Theorem 7.1.

THEOREM 7.2. *Assume* f *satisfies* (NS_1), *except that*

$$f'(0)=-k, \qquad 0<k\leqq 1,$$

holds instead of $f'(0)<-1$. *If* $0<k<1$, *assume that* $f''(0)\geqq 0$. *Then with* $f_*(x)$ *given by* (5.3) *we have*

$$|f_*^2(x)|<|x| \quad \forall x\neq 0.$$

Note that (7.1) and (7.2) hold under the hypotheses of Theorem 7.2.

THEOREM 7.3. *Assume* f *satisfies both* (NS_2) *and* (I). *Then* f *satisfies* (II) *if and only if both*

(i) $|(f^2)'(x)|\leqq 1$, *and*

(ii) $(f^2)''(x)=0$ *if* $(f^2)'(x)=1$

hold whenever

$$f^2(x)=x, \qquad x\in(0,\alpha), \quad f(x)\in(-\beta, 0). \tag{7.4}$$

Equivalently, f *satisfies* (II) *if and only if both* (i) *and* (ii) *hold whenever*

$$f^2(x)=x, \qquad x\in(-\beta, 0), \quad f(x)\in(0,\alpha). \tag{7.5}$$

LEMMA 7.1. *Assume* f *satisfies* (NS_1), *and define* f_*, A_*, *and* B_* *by* (5.3) *and* (5.6). *Then* A_* *and* $-B_*$ *are the unique nonzero fixed points of* f_*^2. *Moreover,* $f_*^2(x)-x$

changes sign at these points, with $|f_*^2(x)|>|x|$ *in* $(-B_*, A_*)-\{0\}$, *and* $|f_*^2(x)|<|x|$ *in* $(-\infty, -B_*)\cup(A_*, \infty)$.

Proof. First observe that a (possibly infinite) quantity $\xi_*>0$ exists such that

$$(f_*^2)'(x)>0 \quad \text{in } [0, \xi_*),$$

$$f_*^2(x)=f_*^2(\xi_*) \quad \text{in } [\xi_*, \infty) \quad \text{if } \xi_*<\infty.$$

This follows easily from the definition of f_*, and we see that $\xi_*=\xi$ if $f_*(\xi)\geqq-\eta$ and $\xi_*=f_*^{-1}(-\eta)<\xi$ if $f_*(\xi)<-\eta$ (these formulas hold even if ξ or η is infinite). Also,

$$Sf_*^2(x)<0 \quad \text{in } [0, \xi_*)$$

by (ii) of Proposition 6.1. By assumption, we have

$$(f_*^2)'(0)>1. \tag{7.6}$$

Now A_* is the smallest positive fixed point of f_*^2 in $(0, \infty)$. If $A_*\geqq\xi_*$, then clearly A_* is the only such fixed point; so suppose that $A_*<\xi_*$. Because A_* is the smallest positive fixed point of f_*^2, we have $f_*^2(x)>x$ for $0<x<A_*$, which implies

$$(f_*^2)'(A_*)\leqq 1. \tag{7.7}$$

By Lemma 6.1 the derivative $(f_*^2)'$ does not attain a minimum in any open subinterval of $(0, \xi_*)$. Using this and the fact that $(f_*^2)'(0)>1$, we conclude from (7.7) that

$$(f_*^2)'(x)<1 \quad \text{for } A_*<x<\xi_*. \tag{7.8}$$

Integrating (7.8) from A_* to u for $A_*<u\leqq\xi_*$, we obtain

$$f_*^2(u)-f_*^2(A_*)=f_*^2(u)-A_*<\int_{A_*}^{u} 1\,dx=u-A_*, \tag{7.9}$$

and (7.9) implies $f_*^2(x)<x$ for $A_*<x\leqq\xi_*$, and hence for all $x>A_*$.

The analysis for $-B_*$ is similar and is left to the reader. □

Proof of Theorem 7.1. As was noted in § 5, the sufficient conditions (5.7) and (5.11) in Propositions 5.1 and 5.2 are also necessary if f_*^2 has a unique positive fixed point. Lemma 7.1 shows that f_*^2 has a unique positive fixed point.

Proof of Theorem 7.2. Obviously f^2 and f_*^2 agree on $[0, \xi_*]$. A simple calculation (see [27, p. 261]) shows that

$$(f_*^2)''(0)=f''(0)[k^2-k]. \tag{7.10}$$

Since we assume that $f''(0)\geqq 0$ if $0<k<1$, (7.10) implies

$$(f_*^2)''(0)\leqq 0 \tag{7.11}$$

for $0<k\leqq 1$. If strict inequality holds in (7.11), the mean value theorem implies that there exists $\delta>0$ such that

$$0<(f_*^2)'(x)<(f_*^2)'(0)=k^2 \quad \text{for } 0<x<\delta. \tag{7.12}$$

If $(f_*^2)''(0)=0$, the negative Schwarzian condition implies

$$(f_*^2)'''(0)<0,$$

and by using Taylor's formula we again see that there exists $\delta>0$ such that (7.12) is satisfied. Lemma 6.1 now implies

$$(f_*^2)'(x)<k^2 \quad \text{for } 0<x<\xi_*, \tag{7.13}$$

for if (7.13) failed for some x, $(f_*^2)'$ would achieve its minimum at an interior point of $(0, x)$. By integrating inequality (7.13) from zero to x for $x \leqq \xi_*$, we easily obtain

$$f_*^2(x) < x \quad \text{for } 0 < x \leqq \xi_*,$$

and hence

$$f_*^2(x) < x \quad \text{for } 0 < x. \tag{7.14}$$

Inequality (7.14) implies that f_*^2 has no negative fixed points y (otherwise $f_*(y)$ would be a positive fixed point of f_*^2), and since $f_*^2(x) > x$ for small negative x, we conclude that $f_*^2(x) > x$ for all $x < 0$. □

LEMMA 7.2. *Assume f satisfies both* (NS_1) *and* (I). *If $a \in (-\beta, \alpha)$ is such that $f(a) \in (-\beta, \alpha)$ and $f^2(a) = a$, then we have in fact $a, f(a) \in [-B, A] \subseteq (-\beta, \alpha)$ where A and B are as in* (I).

Proof. Assume without loss of generality that $a > 0$. The monotonicity of f_* and the fact that $|f(x)| \leqq |f_*(x)|$ in $(-\beta, \alpha)$ imply

$$a = f^2(a) \leqq f_*(f(a)) \leqq f_*^2(a),$$

and from this we have

$$a \leqq A_* \tag{7.15}$$

by Lemma 7.1. On the other hand, the inclusion $f^2([-B, A]) \subset [-B, A]$ (which follows from (I)) and (5.3) imply that $f_*^2(A) \leqq A$, so

$$A_* \leqq A \tag{7.16}$$

by Lemma 7.1. From (7.15) and (7.16) we have $a \in [-B, A]$. The proof that $f(a) \in [-B, A]$ is analogous. □

LEMMA 7.3. *Assume f satisfies both* (PM) *and* (I), *and is differentiable in $(-\beta, \alpha)$ with $f'(x) \neq 0$ there, except at $x = \xi$ and $x = -\eta$ (this is true in particular if f satisfies* (NS_2) *and* (I)). *Then with A as in* (I), *the following hold*:

(i) *The critical points of f^2 in $(0, A)$ are isolated, and $(f^2)'$ changes sign at each such point.*

(ii) *If a point w in the open interval $(0, A)$ is a local maximum of f^2, then it is a global maximum in $[0, A]$*:

$$f^2(w) = \max_{[0,A]} f^2(x). \tag{7.17}$$

(iii) *If f^2 possesses a critical point in the closed interval $[0, A]$ and if w in the closed interval $[0, A]$ is as in* (7.17), *then w is a critical point*:

$$(f^2)'(w) = 0.$$

Proof. This lemma follows directly from several elementary observations based on the shape of the graph of f as in (PM), and the fact that f maps the interval $[-B, A] \subseteq (-\beta, \alpha)$ into itself.

At a critical point $x \in (0, A)$ of f^2 we have

$$(f^2)'(x) = f'(f(x)) f'(x) = 0$$

and so either $x = \xi$ or $f(x) = -\eta$. As $f(x) = -\eta$ for at most two points in $(0, \alpha)$, we conclude that f^2 has at most three critical points in $(0, A)$. Of course, these points are isolated. If either $A \leqq \xi$ or $f(\xi) \geqq -\eta$, then f^2 has at most one critical point in $(0, A)$, and this point, if it exists, is a local maximum. In this case (i), (ii), and (iii) clearly hold, so the lemma is proved.

On the other hand, suppose $A > \xi$ and $f(\xi) < -\eta$. Then we see that $x = \xi$ is a local minimum of f^2, that $f(x) = -\eta$ has either one or two solutions ζ in $(0, A]$ and that they are local maxima for f^2 with the (common, if there are two solutions ζ) value

$$f^2(\zeta) = f(-\eta) = \max_{[-B,0]} f(x) = \max_{[0,A]} f^2(x). \tag{7.18}$$

In particular, (i) and (ii) hold. To prove (iii), we note that if $w \in [0, A]$ satisfies (7.17), then from (7.18) we have $f^2(w) = f(-\eta)$, and hence $f(w) = -\eta$. Thus $(f^2)'(w) = f'(-\eta)f'(w) = 0$ as claimed. □

LEMMA 7.4. *Assume f satisfies the hypotheses of Lemma* 7.3. *Suppose there exists an interval $J = [r, s]$ with $0 \leqq r < s \leqq A$ such that*

$$f^2(J) = J,\ f^2(\partial J) = \partial J,\ \text{where } \partial J = \{r, s\},\ \text{and } J - \partial J \text{ contains a critical point of } f^2. \tag{7.19}$$

Furthermore, assume that it is not true that $f^2(r) = s = f^2(s)$. Then there exists $v \in J - \partial J$ such that $f^2(v) = s$. Also, if $w \in \partial J$ is such that $f^2(w) = s$, then $(f^2)'(w) = 0$.

Proof. Our assumptions imply that (a) $f^2(r) = s$ and $f^2(s) = r$, or (b) $f^2(r) = r$ and $f^2(s) = s$, or (c) $f^2(r) = r = f^2(s)$. In case (a), let $v = \sup\{x < s: (f^2)'(x) = 0\}$. By using (i) of Lemma 7.3, the fact that f^2 achieves its minimum on J at s and the assumption that f^2 has a critical point in (r, s), we see that $r < v < s$, and $(f^2)'(x) < 0$ for $v < x < s$. Lemma 7.3 implies that $(f^2)'(x)$ changes sign at v, so f^2 has a local maximum at v. A similar argument applies in cases (b) and (c) and shows that f^2 always has a critical point v in (r, s) at which f^2 has a local maximum. Note, however, that this argument fails if $f^2(r) = s = f^2(s)$.

Because $f^2(J) \subset J$, we have $f^2(v) \leqq s$; but part (ii) of Lemma 7.3 implies

$$f^2(v) = \max_{[0,A]} f^2(x) \geqq s, \tag{7.20}$$

so we conclude that

$$f^2(v) = s = \max_{[0,A]} f^2(x). \tag{7.21}$$

If there exists $w \in \partial J$ such that $f^2(w) = s$, (7.21) and part (iii) of Lemma 7.3 imply that $(f^2)'(w) = 0$. □

Proof of Theorem 7.3. We establish the last statement of the theorem first, by showing (i) and (ii) both hold for all fixed points of f^2 satisfying (7.4) if and only if they hold for all fixed points of f^2 satisfying (7.5). Indeed, this is an easy consequence of the following two observations. First, if $f^2(x) = x$ then $f^2(y) = y$, where $y = f(x)$, and we have $(f^2)'(x) = (f^2)'(y)$. Second, if $(f^2)'(x) = 1$ for this point, then $(f^2)''(x) = (f^2)''(y)$. Thus, we need only consider fixed points of f^2 satisfying (7.5).

By Lemma 7.2 we may further restrict our attention to fixed points $x \in (0, A]$ (with $f(x) \in [-B, 0)$ holding automatically), where A and B are as in (I). We will therefore prove that for f to satisfy (II) it is necessary and sufficient that each fixed point of f^2 in $(0, A]$ should satisfy both (i) and (ii).

Necessity. Assume that f satisfies (II) for some a and b, but that either (i) or (ii) fails for $x = a$. From (II) we see that $x = a$ is the only fixed point of both f^2 and of f^4 in $(0, A]$, and so we have

$$\begin{aligned} &f^n(x) > x \quad \text{in } (0, a), \\ &f^n(x) < x \quad \text{in } (a, A] \quad \text{if } a < A \end{aligned} \tag{7.22}$$

for $n=2$ and 4, because $(f^n)'(0)>1$ and $f^n(A)\leqq A$. Observing that $f^2(a)=a$ implies $(f^4)'(a)=[(f^2)'(a)]^2\geqq 0$, we conclude from (7.22) that $(f^4)'(a)\leqq 1$; hence $|(f^2)'(a)|\leqq 1$. Thus (i) holds for $x=a$.

We therefore assume that (ii) fails for $x=a$, and so

$$(f^2)'(a)=1 \quad \text{and} \quad (f^2)''(a)\neq 0. \tag{7.23}$$

We now see that in order for (7.22) to hold with $n=2$ it is necessary that

$$(f^2)''(a)>0. \tag{7.24}$$

Furthermore, the fixed point a must be located at the endpoint A of the interval, that is,

$$f^2(A)=A \quad \text{as } a=A.$$

Now observe that the interval $(0, A)$ contains a critical point of f^2. Indeed if this were not so, then $(f^2)'$ would attain a positive minimum in $(0, A)$ because of (7.23) and the fact that $(f^2)'(0)>1$; however, this would contradict Corollary 6.1.

Thus we see that the interval $J=[0, A]$ satisfies the hypotheses of Lemma 7.4. But then we conclude from this result, with $w=A$, that $(f^2)'(A)=0$. This contradicts (7.23), completing the proof of necessity.

Sufficiency. Assume that f satisfies (NS_2) and (I) for some A and B, and that (i) and (ii) in the statement of the theorem hold at each fixed point of f^2 in $(0, A]$. First note that f^2 has at least one fixed point in $(0, A]$; this follows from the inclusion $f^2([\delta, A])\subseteq[\delta, A]$, which is true for sufficiently small $\delta>0$ because $(f^2)'(0)>1$. Choose any such fixed point a, that is,

$$f^2(a)=a\in(0, A],$$

and consider its domain of attraction W in $[0, A]$ defined by

$$W=\{x\in[0, A]\,|\,f^{2n}(x)\to a \text{ as } n\to\infty\}.$$

Clearly $a\in W$ and $0\notin W$. By using Lemma 6.1 with f^2 in place of f we see that the set W is relatively open in $[0, A]$. Let $I\subseteq W$ denote the maximal connected component of W containing $x=a$; thus I is an interval of the form

$$I=(r, s), \quad \text{or else} \quad I=(r, A]$$

where in either case the quantities r and s satisfy

$$0\leqq r<a, \quad a<s\leqq A \quad \text{if } a<A.$$

Because $f(W)\subseteq W$, $f(I)$ is a connected subset of W containing a, the maximality of the connected component implies

$$f(I)\subseteq I. \tag{7.25}$$

Continuity implies that $f^2(\bar{I})\subset\bar{I}$. However, if $I=(r, s)$, we must have that

$$f^2(r), f^2(s)\in\{r, s\}; \tag{7.26}$$

otherwise r or s would be in I. If $I=(r, A]$, the same reasoning implies

$$f^2(r)=r. \tag{7.27}$$

Now observe that neither the point r nor s (if $I=(r, s)$) can be a nonzero fixed point of f^2. By Lemma 6.1 we know that each fixed point of f^2 in $(0, A]$ attracts iterates

$f^{2n}(x)$ of all nearby points x. However, we know that those points $x \in I$ near r or s satisfy $f^{2n}(x) \to a$ instead. Therefore, if $I = (r, s)$, we must have

$$f^2(s) = r,$$

and we must have $f^2(r) = s$ unless $r = 0$. If $I = (r, A]$, we must have $f^2(r) = r$, and we have just seen that this implies $r = 0$. Thus, if $I = (r, A]$, the domain of attraction of the fixed point a of f^2 is the entire interval $(0, A]$, so condition (II) holds. Therefore for the remainder of the proof we assume $I = (r, s)$.

Because $f^2(s) = r$ and $s > 0$ we have $r > 0$, so the previous remarks imply that $f^2(r) = s$. Define $g = f^4$, so g maps $[r, s]$ into itself, g has negative Schwarzian derivative on $[r, s]$, and r, a, and s are fixed points of g. Note that

$$f'(x) = (f^2)'(f^2(x))(f^2)'(x), \tag{7.28}$$

so

$$0 \leqq g'(a) = ((f^2)'(a))^2 \leqq 1. \tag{7.29}$$

Lemma 2.6 of [27] proves that if g is any continuous function on an interval $[r, s]$ and if $g(r) = r$, $g(s) = s$, g is C^3 on (r, s) and $g'(x) \neq 0$ and $(Sg)(x) < 0$ for $x \in (r, s)$, then $g'(a) < 1$ if $a \in (r, s)$ and $g(a) = a$. (Note that the proof of Lemma 2.6 in [27] requires only that g' not vanish on (r, s), although the result is stated slightly less generally.) Thus by Lemma 2.6 of [27] and (7.29) we obtain a contradiction unless $g'(x_0) = 0$ for some $x_0 \in (r, s)$. Because $f^2(I) \subset I$, (7.28) implies f^2 has a critical point in I. Lemma 7.4 now implies there exists $v \in I$ such that $f^2(v) = s$. Since $\lim_{n \to \infty} f^{2n}(x) = a$ for any $x \in I$ and $f^{2n}(v) = r$ or s, we have a contradiction, and the proof is complete. □

If $g(x, \theta)$ is defined for (x, θ) near $(0, \theta_*)$, and if $g(0, \theta) = 0$ and $\partial g(0, \theta_*)/\partial x = -1$, it is natural to ask whether the map $x \to g(x, \theta)$ satisfies (III) at zero for some interval $(\theta_* - \delta, \theta_*)$ or $(\theta_*, \theta_* + \delta)$, $\delta > 0$. This question was answered by Allwright in [1]. He assumed that $x \to g(x, \theta)$ has negative Schwarzian derivative for all x, but his proof only requires that $x \to g(x, \theta)$ have negative Schwarzian derivative for x near zero and θ near θ_*. Thus we obtain the following result, which may also be obtained as a simple consequence of Theorem 7.2.

COROLLARY 7.1 (see Allwright [1]). *Assume $g(x, \theta)$ is defined and continuous for $|x| < \delta_1$ and $|\theta - \theta_*| < \delta_2$. In addition suppose g is C^3 in the x-variable and g has a negative Schwarzian derivative (with respect to the x-variable) at $x = 0$ and $\theta = \theta_*$. Assume $g(0, \theta) = 0$ for $|\theta - \theta_*| < \delta_2$ and $g'(0, \theta_*) = \partial g(0, \theta_*)/\partial x = -1$. Finally, assume $\partial^2 g/\partial\theta\partial x$ is defined and continuous on the domain of g and $\partial^2 g(0, \theta_*)/\partial\theta\partial x \neq 0$. Then there exists $\delta > 0$ such that the map $x \to g(x, \theta)$ satisfies (III) for $\theta_* < \theta < \theta_* + \delta$ if $\partial^2 g(0, \theta_*)/\partial\theta\partial x < 0$, while the map $x \to g(x, \theta)$ satisfies (III) for $\theta_* - \delta < \theta < \theta_*$ if $\partial^2 g(0, \theta_*)/\partial\theta\partial x > 0$.*

Of course what Allwright really shows is that fixed points of period 2 of the map $x \to g(x, \theta)$ are bifurcating at $\theta = \theta_*$ from the trivial fixed point $x = 0$. The negative Schwarzian condition at $x = 0$ ensures that the bifurcation is such that (III) is satisfied. In fact, the negative Schwarzian condition is also essentially necessary for (III) to be satisfied locally. The following proposition indicates the sense in which necessity is meant. Since the proposition follows by standard arguments in local bifurcation theory, the proof is omitted.

PROPOSITION 7.1. *Assume $g(x, \theta)$ is a C^4 function defined on an open neighborhood of $(0, \theta_*)$. Assume $g(0, \theta) = 0$ for $(0, \theta)$ in the domain of g, $\partial g(0, \theta_*)/\partial x = -1$, and $\partial^2 g(0, \theta_*)/\partial\theta\partial x = \eta_1 \neq 0$. Suppose the Schwarzian derivative of g (with respect to the x-variable) at $x = 0$ and $\theta = \theta_*$ is nonzero and denote this Schwarzian derivative by η_2.*

Then there exist positive numbers ε and δ and continuous functions $\xi_+(\theta)$ and $\xi_-(\theta)$ that are defined for $\theta_* \leqq \theta \leqq \theta_* + \delta$ if $\eta_1\eta_2 > 0$ and for $\theta_* - \delta \leqq \theta \leqq \theta_*$ if $\eta_1\eta_2 < 0$ and that have the following properties:

(1) $g(\xi_+(\theta), \theta) = \xi_-(\theta)$ *and* $g(\xi_-(\theta), \theta) = \xi_+(\theta)$,

(2) $\xi_+(\theta_*) = 0 = \xi_-(\theta_*)$ *and the range of* ξ_+ *is in* $[0, \varepsilon]$ *and the range of* ξ_- *is in* $[-\varepsilon, 0]$, *and*

(3) *if* $g(g(x, \theta), \theta) = x$ *for some* (x, θ) *with* $0 < |x| < \varepsilon$ *and* $|\theta - \theta_*| \leqq \delta$, *then* θ *must be in the domain of* ξ_+ *and* $x = \xi_+(\theta)$ *or* $x = \xi_-(\theta)$.

Note that if the Schwarzian derivative is positive, the map $x \to g(x, \theta)$ takes $[\xi_-(\theta), \xi_+(\theta)]$ into itself, but $\partial g(0, \theta)/\partial x > -1$ for θ in the domain of ξ_+.

We need one more theorem for our applications in § 9. Roughly speaking, our next result asserts that for functions with negative Schwarzian derivative, (II) fails before (I).

THEOREM 7.4. *Assume f satisfies* (PM) *and α or β is finite (α, β, ξ, and η are as in the definition of* (PM)). *Suppose f is C^3 on $[-\beta, \alpha]$, $f'(x) \neq 0$ for $x \neq -\eta$ and $x \neq \xi$, and the Schwarzian derivative $Sf(x)$ is negative on $(-\beta, \alpha)$ for $x \neq -\eta, \xi$. If $f([-\beta, \alpha]) \subset [-\beta, \alpha]$ and $f_*^2(\alpha) = \alpha$ or $f_*^2(-\beta) = -\beta$ (f_* as in* (5.3)), *then there exists $\gamma \in (-\beta, \alpha)$, $\gamma \neq 0$, such that $(f^2)(\gamma) = \gamma$ and $(f^2)'(\gamma) < -1$.*

Proof. Assume for definiteness that $f_*^2(\alpha) = \alpha$. We assume that the theorem is false, so $(f^2)'(\gamma) \geqq -1$ for every nonzero $\gamma \in [-\beta, \alpha]$ such that $f^2(\gamma) = \gamma$, and we try to obtain a contradiction. Recall that Lemma 6.1 implies that if $f^2(\gamma) = \gamma$ and $-1 \leqq (f^2)'(\gamma) < 1$, then γ is a "locally stable fixed point of f^2" in the sense that there exists $\delta > 0$ such that $\lim_{n\to\infty} f^{2n}(x) = \gamma$ for all x such that $|x - \gamma| < \delta$.

There are two cases to consider, each corresponding to a different qualitative appearance of f^2.

Case 1. Assume that $-\eta < f(\xi) = f_*(\alpha)$. In this case we have $f_*^2(\alpha) = f^2(\xi) = \alpha$, and we can easily verify that $(f^2)'(x) > 0$ for $0 \leqq x \leqq \xi$ and $(f^2)'(x) < 0$ for $\xi < x < \alpha$. For notational convenience, define $\xi = \xi_2$ in Case 1.

Case 2. Assume that $f(\xi) = f_*(\alpha) \leqq -\eta$. In this case we have $f_*^2(\alpha) = f(-\eta) = \alpha$. Furthermore, there exist a unique number ξ_1, $0 < \xi_1 \leqq \xi$ and a unique number ξ_2, $\xi \leqq \xi_2 < \alpha$ such that $f(\xi_1) = f(\xi_2) = -\eta$. Using this information, we can easily check that $(f^2)'(x) > 0$ for $0 \leqq x < \xi_1$, $(f^2)'(x) < 0$ for $\xi_1 < x < \xi$, $(f^2)'(x) > 0$ for $\xi < x < \xi_2$ and $(f^2)'(x) < 0$ for $\xi_2 < x < \alpha$.

It follows that (in Case 1 or Case 2), $f^2(\xi_2) = \alpha$ and $(f^2)'(x) < 0$ for $\xi_2 < x < \alpha$. Because $f^2(\alpha) = \alpha$, the intermediate value theorem implies that there is a unique number γ, $\xi_2 < \gamma < \alpha$, such that $f^2(\gamma) = \gamma$. Our assumptions imply

$$-1 \leqq (f^2)'(\gamma) \leqq 0,$$

so our previous remarks imply that γ is a locally stable fixed point of f^2.

Just as in the proof of Theorem 7.3, let $U = \{x \in [0, \alpha] : f^{2n}(x) \to \gamma\}$, so U is an open set, and let U_1 be the maximal connected component of U containing γ, so U_1 is also an open set and $f^2(U_1) \subset U_1$. (Note that $0 \notin U$ and $\alpha \notin U$.) Since U_1 is connected, we can write $U_1 = (r, s)$. If $r < \xi_2$, we have $\xi_2 \in U_1$, so $f^4(\xi_2) = 0 \in U_1$, a contradiction. Thus we must have $\xi_2 \leqq r$. If $s = \alpha$, we obtain $0 = f^2(\alpha) \in U_1$ (because $f^2(U_1) \subset U_1$), and this is impossible because $r \geqq \xi_2$. Thus we must have $s < \alpha$. Just as in the proof of Theorem 7.3, we must have $f^2(r) \in \{r, s\}$ and $f^2(s) \in \{r, s\}$.

There are several possibilities to consider. If $f^2(r) = r$ or $f^2(s) = s$, we contradict the fact that γ is the unique fixed point of f^2 in the interval $[\xi_2, \alpha]$. The only other possibility is that $f^2(r) = s$ and $f^2(s) = r$. If we write $g = f^4$, we know that g has a negative Schwarzian derivative on $[r, s]$ and that r, s and γ are fixed points of g. If

$g'(x) \neq 0$ for $r < x < s$, Lemma 2.6 of [27] implies that $g'(\gamma) > 1$, which is a contradiction. Thus there must exist $x_0 \in (r, s)$ such that $g'(x_0) = 0$. By using the chain rule we see that

$$x_0 = \xi \quad \text{or } f^2(x_0) = \xi \quad \text{or } f(x_0) = -\eta \quad \text{or } f^3(x_0) = -\eta.$$

Because $f^2(x_0) \in (r, s)$, $x_0 \in (r, s)$, and $r \geqq \xi_2$, the only possibility is that $f(x_0) = -\eta$ or $f^3(x_0) = -\eta$. In particular, we must be in Case 2 and have $f(\xi) \leqq -\eta$ and $f(-\eta) = \alpha$. But then we again obtain a contradiction: either $\alpha = f^2(x_0) \in (r, s)$ or $\alpha = f^4(x_0) \in (r, s)$. Since we have obtained a contradiction in all cases, the theorem is proved. □

Remark 7.1. Note that we have actually proved somewhat more than is claimed. If f is as in Theorem 7.4 and $f_*^2(\alpha) = \alpha$ and ξ_2 is as defined in the proof, an examination of the previous argument shows that there exists γ with $\xi_2 < \gamma < \alpha$ such that $f^2(\gamma) = \gamma$ and $(f^2)'(\gamma) < -1$. An analogous statement is true if $f_*^2(-\beta) = -\beta$.

In fact, the same kinds of arguments used in Theorem 7.4 allow a much more detailed picture of the fixed points of $g = f^2$. If f is as in Theorem 7.4 and $f_*^2(\alpha) = \alpha$, and ξ_1 and ξ_2 are as in the proof of Theorem 7.4 (so that $\xi_1 = \xi_2 = \xi$ if $f(\xi) \geqq -\eta$), then we can prove g has no nonzero fixed points on $[0, \xi_1]$. If $g(\xi) \leqq \xi$, then g has unique fixed points γ_1 in $[\xi_1, \xi]$ and γ_2 in $(\xi, \xi_2]$, and $g'(\gamma_2) > 1$. If $g(\xi) > \xi$, then g has no fixed points in $[\xi_1, \xi]$ and g either has zero, 1, or 2 fixed points in $[\xi, \xi_2]$. If g has exactly one fixed point γ_1 in $[\xi, \xi_2]$, then $g'(\gamma_1) = 1$ and $g''(\gamma_1) > 0$. If g has exactly two fixed points $\gamma_1 < \gamma_2$ in $[\xi, \xi_2]$, then $0 < g'(\gamma_1) < 1$ and $g'(\gamma_2) > 1$. Because it is very long, we omit the proof.

COROLLARY 7.2. *Suppose f is as in Theorem* 7.4 *and $f_n : \mathbb{R} \to \mathbb{R}$, $n \geqq 1$, is a sequence of C^1 functions such that $f_n(x) \to f(x)$ and $f_n'(x) \to f'(x)$ uniformly on compact intervals. Assume f_n satisfies condition* (0) *and positive numbers A_n and B_n exist such that $f_n([-B_n, A_n]) \subset [-B_n, A_n]$, $xf_n(x) < 0$ for all $x \in [-B_n, A_n] - \{0\}$, and $A_n \to \alpha$ and $B_n \to \beta$ as $n \to \infty$ (α and β are as in the definition of* (PM) *for f). If $f_*^2(\alpha) = \alpha$ or $f_*^2(\beta) = \beta$, then there exists $\gamma \in (-\beta, \alpha)$ such that $f^2(\gamma) = \gamma$ and $(f^2)'(\gamma) < -1$, and there exists a sequence $(\gamma_n) \to \gamma$, defined for n sufficiently large, such that $\gamma_n \in (-B_n, A_n)$, $f_n^2(\gamma_n) = \gamma_n$ and $(f_n^2)'(\gamma_n) < -1$.*

Proof. This follows immediately from Theorem 7.4 and elementary calculus arguments. □

8. The Schwarzian derivative of f_k. To apply the theory of § 7 to the functions f_k, we must first show $Sf_k(x) < 0$ for the appropriate ranges of x. As the Schwarzian derivative is invariant under translation, it is sufficient to work directly with the functions f_k rather than with the corresponding normalized functions

$$g_k(x) = f_k(x + x_0) - f_k(x_0). \tag{8.1}$$

PROPOSITION 8.1. *For the functions f_k we have*

$$Sf_k(x) < 0 \quad \text{whenever } f_k'(x) \neq 0 \tag{8.2}$$

for the ranges of parameters and values of x in Table 2. *Also, hypothesis* (NS_1) *or* (NS_2) *holds at the fixed point x_0 for all $\mu > \mu_0$ as indicated in Table* 2, *where x_0 and μ_0 are as in Table* 1.

The data of Table 2 are sufficient but not necessary for the Schwarzian derivative of f_k to be negative, or for (NS_1) or (NS_2) to hold. For example, we have not ruled out the possibility that in at least part of the range $0 < \nu < 1$ the condition (NS_2) might hold for f_4 or f_5, rather than the weaker condition (NS_1).

Proof. Rather than prove this result by direct (but lengthy) calculation of the Schwarzian derivatives, we use the results of § 6 to simplify our work.

TABLE 2

Ranges where the Schwarzian derivative of f_k is negative (provided $f_k'(x) \neq 0$), and where (NS_1) *or* (NS_2) *holds at the fixed point x_0 for all $\mu > \mu_0$. The values of x_0 and μ_0 are as in Table* 1.

k	Range where the Schwarzian derivative is negative (where (8.2) holds)	Hypothesis (NS_1) or (NS_2) holding at x_0 for all $\mu > \mu_0$
1	all $x \in \mathbb{R}$, $\mu \in \mathbb{R}$	(NS_2)
2	all $x \in \mathbb{R}$ when $\mu \geqq 0$	(NS_2)
3	all $x \in \mathbb{R}$, $\mu, \theta \in \mathbb{R}$	(NS_2) when $0 \leqq \theta < \pi/2$
4	$x > 0$ when $\nu \geqq 1$	(NS_2) when $\nu \geqq 1$ or $\nu = 0$
	$x > \nu$ when $0 \leqq \nu < 1$	(NS_1) when $0 < \nu < 1$
5	$x > 0$ when $\nu \geqq 1$ and $\lambda > \nu + 1$	(NS_2) when $\nu \geqq 1$ and $\lambda > \nu + 1$, or when $\nu = 0$ and $\lambda > 1$
	$x > \left(\frac{\nu}{\lambda - \nu}\right)^{1/\lambda}$ when $0 \leqq \nu < 1$ and $\lambda > \nu + 1$	(NS_1) when $0 < \nu < 1$ and $\lambda > \nu + 1$

The cases $k = 1$, 2, and 3 follow immediately from Proposition 6.2. (An easy alternate proof in the case of f_3 is to note that $Sf_3(x) \leqq f_3'''(x)/f_3'(x) = -1$ wherever $f_3'(x) \neq 0$.) Essentially the same argument as in the proof of Proposition 6.2 also works for f_4, even though this is not an entire function when ν is not an integer. Assuming $\mu \neq 0$, for $x > 0$ and $x \neq \nu$ we have

$$\frac{d^2}{dx^2} \log |f_4'(x)| = -\frac{\nu - 1}{x^2} - \frac{1}{(x - \nu)^2}, \tag{8.3}$$

which is negative if $\nu \geqq 1$. If $0 < \nu < 1$ and $x > \nu$ then (8.3) is bounded above by $-(\nu - 1)/x^2 - 1/x^2 = -\nu/x^2$, and hence is again negative. When $\nu = 0$, a direct calculation of the Schwarzian derivative shows that $Sf_4(x) = -\frac{1}{2}$ for $x > 0$. The claims of the proposition now follow directly. In particular, we see that (NS_1) holds for f_4 when $0 < \nu < 1$ and $\mu > \mu_0$, when we note that f_4 achieves its maximum at $x = \nu$ and this critical point lies to the left of the fixed point x_0.

The calculations for f_5 are somewhat more involved, but some simplification can be achieved by considering the reciprocal of this function. Setting $m(x) = \mu/x$ where $\mu \neq 0$, from (iii) of Proposition 6.1 we have that for $x > 0$ and $f_5'(x) \neq 0$

$$Sf_5(x) = Sh(x)$$

where

$$h(x) = m(f_5(x)) = x^{-\nu} + x^{\lambda - \nu}.$$

Further calculation yields

$$Sh(x) = -\frac{Py^2 + Qy + R}{x^2[(\lambda - \nu)y - \nu]^2}$$

where

$$P = \tfrac{1}{2}(\lambda - \nu)^2[(\lambda - \nu)^2 - 1],$$

$$Q = \nu(\lambda - \nu)[(\lambda - \nu)^2 + 3\nu(\lambda - \nu) + \nu^2 + 1],$$

$$R = \tfrac{1}{2}\nu^2(\nu^2 - 1),$$

$$y = x^{\lambda}, \qquad y \neq \frac{\nu}{\lambda - \nu}.$$

If $\nu \geqq 0$ and $\lambda > \nu+1$ as in Table 1 then $P>0$ and $Q>0$. If in addition either $\nu \geqq 1$ or $\nu = 0$, then $R \geqq 0$, and so $Sf_5(x)<0$ for $x>0$. In this case f_5 satisfies (NS_2) when $\mu > \mu_0$, as claimed. On the other hand, if $\lambda > \nu+1$ but $0<\nu<1$, then a calculation reveals $Py^2+Qy+R>0$ at $y=\nu/(\lambda-\nu)$. Thus,

$$Py^2+Qy+R>0 \quad \text{for } y \geqq \frac{\nu}{\lambda-\nu}$$

because $P>0$ and $Q>0$; hence $Sf_5(x)<0$ for $x>(\nu/(\lambda-\nu))^{1/\lambda}$. Again, as f_5 achieves its maximum at $x=(\nu/(\lambda-\nu))^{1/\lambda}$ and this point lies to the left of x_0 when $\mu>\mu_0$, it follows that f_5 satisfies (NS_1), as claimed. □

9. Applications of the general theory to the function f_k. We will now use the results of § 7 to determine ranges of parameters for which hypothesis (I), (II), or (III) holds for f_k at a fixed point x_0 of Table 1. In this connection it will first be useful to make a few general remarks.

Suppose that $f(x)$ is a continuous function with fixed point x_0 and assume that the function $g(x)$ defined by

$$f(x+x_0)-x_0=g(x)$$

satisfies (PM). Recall that f satisfies (I), (II), or (III) at x_0 if and only if g satisfies the corresponding hypothesis at zero. If α, β, ξ, and η are the quantities in (PM) for g, then define $x_1=x_0-\beta$, $x_2=x_0-\eta$, $x_3=x_0+\xi$, and $x_4=x_0+\alpha$. We are considering the function $f(x)$ on the interval $[x_1, x_4]$, and $f(x)>x_0$ for $x_1<x<x_0$, $f(x_1)=x_0$ if $x_1>-\infty$, $f(x)<x_0$ for $x_0<x<x_4$, and $f(x_4)=x_0$ if $x_4<\infty$. Furthermore, $[x_2, x_3]$ is the maximum interval containing x_0 on which f is monotone decreasing and f is monotone increasing on $[x_1, x_2]$ and $[x_3, x_4]$. Upon defining the function $f_*:\mathbb{R}\to\mathbb{R}$ by

$$f_*(x)=\begin{cases} f(x_2) & \text{in } (-\infty, x_2] \quad \text{if } x_2>-\infty, \\ f(x) & \text{in } [x_2, x_3], \\ f(x_3) & \text{in } [x_3, \infty) \quad \text{if } x_3<\infty, \end{cases}$$

we see that (7.1) and (7.2) of Theorem 7.2 for g become

$$\begin{aligned} &(f_*^2)(x_4)<x_4 \quad \text{if } x_4<\infty, \\ &(f_*^2)(x_1)>x_1 \quad \text{if } x_1>-\infty, \end{aligned} \tag{9.1}$$

and

$$\begin{aligned} &f^2(x_3)\leqq x_3 \quad \text{if } x_3<\infty, \\ &f_*^2(x_2)\geqq x_2 \quad \text{if } x_2>-\infty. \end{aligned} \tag{9.2}$$

Thus if g satisfies (PM) and f is C^3 with negative Schwarzian derivative on (x_2, x_3), then f satisfies hypothesis (I) at x_0 if and only if the inequalities (9.1) hold, and f satisfies (III) if and only if (9.2) holds.

Now suppose that f is a continuous function with fixed point x_0, that g satisfies (PM), and x_1, x_2, x_3, and x_4 are as defined above. Suppose that φ is a C^1 function defined on an interval (y_1, y_4), that $\varphi'(y)>0$ for $y_1<y<y_4$, and that $\varphi(y_1)=x_1$ and $\varphi(y_4)=x_4$. If $x_1=-\infty$, then assume for convenience that $y_1=-\infty$, and similarly if $x_4=\infty$, then assume $y_4=\infty$. It is easy to check that f satisfies hypothesis (I), (II), or (III) at x_0 if and only if $\varphi^{-1}f\varphi$ satisfies the corresponding hypothesis at $y_0=\varphi^{-1}(x_0)$. Furthermore, writing $h=\varphi^{-1}f\varphi$, we easily verify that

$$h_*=\varphi^{-1}f_*\varphi. \tag{9.3}$$

If f is C^3 on $[x_2, x_3]$ and $Sf(x) < 0$ for $x_2 < x < x_3$, it follows from (9.1)-(9.3) and the above remarks that $h = \varphi^{-1} f \varphi$ satisfies (I) at y_0 if and only if

$$(h_*^2)(y_4) < y_4 \quad \text{and} \quad (h_*^2)(y_1) > y_1, \tag{9.4}$$

where $y_j = \phi^{-1}(x_j)$ for $0 \leqq j \leqq 4$. Similarly, h satisfies (III) at y_0 if and only if

$$(h_*^2)(y_3) \leqq y_3 \quad \text{and} \quad (h_*^2)(y_1) \geqq y_1. \tag{9.5}$$

Note that h need not have negative Schwarzian derivative on (y_2, y_3).

The above observation sometimes simplifies calculations, since it may be easier to work with h and h_* than with f and f_*.

We now begin the analyses of the functions f_k. Consider first the function f_1. We easily see that with x_0 as in Table 1 and $\mu > \mu_0 = \frac{3}{4}$ we have $\alpha + x_0 = \xi + x_0 = \infty$, $-\beta + x_0 = -x_0$, and $-\eta + x_0 = 0$, so

$$f_{1*}(x) = \begin{cases} \mu & \text{in } (-\infty, 0], \\ \mu - x^2 & \text{in } (0, \infty). \end{cases}$$

We have $f_{1*}^2(-\eta + x_0) \geqq -\eta + x_0$ if and only if $\mu - \mu^2 \geqq 0$, and $f_{1*}^2(-\beta + x_0) \geqq -\beta + x_0$ if and only if $\mu - \mu^2 \geqq -x_0$, that is,

$$2\mu^2 - 2\mu + 1 < \sqrt{4\mu + 1}. \tag{9.6}$$

Inequality (9.6) is equivalent to

$$\mu^3 - 2\mu^2 + 2\mu - 2 < 0,$$

as a short calculation shows; and this in turn is equivalent to

$$\mu < \mu_* \cong 1.5437$$

where μ_* is the unique real root of $\mu^3 - 2\mu^2 + 2\mu - 2 = 0$. By Theorem 7.1 we conclude from these calculations, and the data of Table 1, that (I) holds at x_0 if and only if $\frac{3}{4} < \mu < \mu_*$, and that (III) holds if and only if $\frac{3}{4} < \mu \leqq 1$.

To determine those values of μ between $\mu = \frac{3}{4}$ and $\mu = \mu_*$ at which (II) holds, we must consider points of period 2 for the map f_1 in the interval $(-\beta + x_0, \alpha + x_0) = (-x_0, \infty)$. Assuming that $\frac{3}{4} < \mu < \mu_*$, we consider points x_1 and x_2 satisfying

$$f_1(x_1) = x_2 \quad \text{and} \quad f_1(x_2) = x_1 \tag{9.7}$$

and lying on either side of x_0 in the above interval:

$$-x_0 < x_1 < x_0 < x_2. \tag{9.8}$$

As noted earlier such points do exist; in fact, in the closed interval $[-B + x_0, A + x_0] \subseteq (-\beta + x_0, \alpha + x_0)$. Writing out the equations in (9.7) gives, after some manipulation, that $x_1 + x_2 = 1$ and

$$\mu = 1 - x_1 x_2.$$

Further, the derivative of f_1^2 at either of these points is

$$(f_1^2)'(x_1) = (f_1^2)'(x_2) = f_1'(x_1) f_1'(x_2) = 4x_1 x_2 = 4(1 - \mu).$$

In the range $\frac{3}{4} < \mu \leqq \frac{5}{4}$ we therefore have $-1 \leqq (f_1^2)'(x_1) < 1$, so (II) holds by Theorem 7.3. Theorem 7.3 also implies the uniqueness of solutions of (9.7) and (9.8) for $\frac{3}{4} < \mu \leqq \frac{5}{4}$. Of course, since $x_1 + x_2 = 1$, we can also solve for x_1 and x_2 and directly obtain uniqueness. If, on the other hand, $\frac{5}{4} < \mu < \mu_*$, the same calculations show $|(f_1^2)'(x_j)| > 1$, and (II) does not hold. Table 3 summarizes our results for f_1 by indicating the parameter ranges where (I), (II), or (III) holds.

The situation for f_2 is very similar to that for f_1. In particular the same sort of analysis as above yields intervals of the parameter μ in which various hypotheses hold. Thus we easily find that f_2 satisfies (I) for $1<\mu<3\sqrt{3}/2$ and (III) for $1<\mu<\frac{3}{2}$. If $f_2(x_1)=x_2$ and $f_2(x_2)=x_1$, where $-\sqrt{\mu}<x_2<0<x_1<\sqrt{\mu}$, then $x_1f_2(x_1)-x_2f_2(x_2)=0$, from which we derive $\mu=x_1^2+x_2^2$, if $x_1\neq -x_2$, or $x_1=-x_2$. The equation

$$f_2(x_1)-x_2-f_2(x_2)+x_1=0$$

implies

$$x_1^2+x_1x_2+x_2^2=\mu-1,$$

so if $x_1\neq -x_2$ we have $x_1^2+x_2^2=\mu$ and

$$x_1x_2=-1. \tag{9.9}$$

Now if $x_2=-x_1^{-1}$, a simple calculation shows that the defining equations for x_2 and x_1 are equivalent to the single equation

$$x_1^4-\mu x_1^2=-1. \tag{9.10}$$

The quadratic equation (9.10) has a real, positive solution if and only if $\mu\geqq 2$. Therefore, if $1<\mu<2$, we must have $x_2=-x_1$. Since f_2 is odd, the defining equations for x_1 and x_2 reduce to the single equation

$$f_2(x_1)=x_1^3-\mu x_1=-x_1,$$

so we obtain

$$x_1=\sqrt{\mu-1} \quad \text{and} \quad x_2=-\sqrt{\mu-1}. \tag{9.11}$$

Substituting (9.11) into the following equation, we obtain

$$\begin{aligned}(f_2^2)'(x_1)=f_2'(x_2)f_2'(x_1)&=(3x_2^2-\mu)(3x_1^2-\mu)\\&=(2\mu-3)^2.\end{aligned} \tag{9.12}$$

Equation (9.12) implies that $|(f_2^2)'(x_j)|<1$ if $1<\mu<2$, so (II) holds for $1<\mu<2$. On the other hand, if $\mu>2$ and we take $x_2=-x_1$, Theorem 7.3 and (9.12) imply that (II) is not satisfied.

Finally, if $\mu=2$, a direct calculation using the above information shows that $x_1=1$ and $x_2=-1$ are the only nonzero fixed points of f_2^2 and

$$(f_2^3)''(x_1)=0,$$

so Theorem 7.3 again implies that (II) is satisfied. All of this information is summarized in Table 3.

TABLE 3

$f_k(x)$ satisfies (I) *at x_0 if and only if $\mu_0<\mu<\mu_1$, $f_k(x)$ satisfies* (II) *if and only if $\mu_0<\mu\leqq\mu_2$, and $f_k(x)$ satisfies* (III) *at x_0 if and only if $\mu_0<\mu\leqq\mu_3$.*

k	x_0	μ_0	μ_1	μ_2	μ_3
1	$(-1+\sqrt{4\mu+1})/2$	$\frac{3}{4}$	$\mu_1^3-2\mu_1^2+2\mu_1-2=0$ $\mu_1\cong 1.5437$	$\frac{5}{4}$	1
2	0	1	$3\sqrt{3}/2$	2	$\frac{3}{2}$

For the map f_3 with fixed point $x_0 = 0$ and parameter θ chosen in the range $0 \leqq \theta < \pi/2$ established earlier, we again find intervals of μ in which (I) and (III) hold. The range where (II) holds, however, is still not clear. Recall the critical value $\mu_0 = \mu_0(\theta) = 1/\cos\theta$ from Table 1.

THEOREM 9.1. *There exist continuous functions*

$$\rho_3, \tau_3 : \left[0, \frac{\pi}{2}\right) \to (0, \infty)$$

satisfying

$$\mu_0(\theta) = \frac{1}{\cos\theta} < \rho_3(\theta) < \tau_3(\theta)$$

such that if $0 \leqq \theta \leqq \pi/2$, *then* (I) *holds for* f_3 *at* $x_0 = 0$ *if and only if* $\mu_0(\theta) < \mu < \tau_3(\theta)$, *and* (III) *holds if and only if* $\mu_0(\theta) < \mu \leqq \rho_3(\theta)$.

Motivated by the results for f_1 and f_2, we might expect the existence of a third function σ_3 satisfying $\rho_3(\theta) < \sigma_3(\theta) < \tau_3(\theta)$ and such that (II) holds if and only if $\mu_0(\theta) < \mu \leqq \sigma_3(\theta)$. We believe this to be the case, but we have not pursued this question here. However, we can easily prove by an implicit function theorem argument that condition (II) holds for $\mu_0(\theta) < \mu < \rho_3(\theta) + \delta_3(\theta)$ for some sufficiently small $\delta_3(\theta) > 0$: the period 2 points $\{x_1 x_2\}$ of f_3 for $\mu = \rho_3(\theta)$ are "super-stable" (that is, $(f_3^2)'(x_1) = (f_3^2)'(x_2) = 0$) and so must persist for μ slightly larger than $\rho_3(\theta)$. On the other hand, Corollary 7.2 implies that there exists $\delta_4(\theta) > 0$ such that for $\tau_3(\theta) - \delta_4(\theta) < \mu < \tau_3(\theta)$ the function f_3^2 has a fixed point γ (in the relevant interval) such that $(f_3^2)'(\gamma) < -1$, and this implies (II) fails for $\tau_3(\theta) - \delta_4(\theta) < \mu < \tau_3(\theta)$. By using Remark 7.1 we can also show that, for θ near zero and μ near $\tau_3(\theta)$, f_3^2 has a second fixed point $\bar{\gamma}$ for which $(f_3^2)'(\bar{\gamma}) > 1$.

Proof of Theorem 9.1. Assuming $\mu > 0$ and $0 \leqq \theta < \pi/2$, we note the following values:

$$\alpha = \pi - \theta \leqq \beta = \pi + \theta, \qquad \xi = \frac{\pi}{2} - \theta \leqq \eta = \frac{\pi}{2} + \theta$$

for f_3 in condition (PM). We also note the following formulas:

$$f_{3*}(\alpha) = f_{3*}(\xi) = -\mu(1 - \sin\theta),$$
$$f_{3*}(-\beta) = f_{3*}(-\eta) = \mu(1 + \sin\theta).$$

In order to use Theorem 7.1 for determining when (I) or (III) holds, we must calculate both $(f_{3*}^2)(\alpha) = (f_{3*}^2)(\xi)$ and $(f_{3*}^2)(-\beta) = (f_{3*}^2)(-\eta)$, and examine (7.1) or (7.2). In fact, we claim

$$\begin{aligned} &f_{3*}^2(\alpha) < \alpha \quad \text{implies } f_{3*}^2(-\beta) > -\beta, \\ &f_{3*}^2(\xi) \leqq \xi \quad \text{implies } f_{3*}^2(-\eta) \geqq -\eta. \end{aligned} \tag{9.13}$$

Thus, we need only verify the inequality $f_{3*}^2(\alpha) < \alpha$ to conclude (I), and $f_{3*}^2(\xi) \leqq \xi$ to conclude (III).

We prove only the first implication (9.13), as the proof of the other is similar. Suppose

$$f_{3*}^2(-\beta) \leqq -\beta. \tag{9.14}$$

Then as f_{3*} achieves its minimum at $x = \xi$, we have

$$f_{3*}(\xi) \leqq -\beta \tag{9.15}$$

from (9.14), and so

$$f_{3*}^2(\xi) = f_{3*}(-\beta) \tag{9.16}$$

as f_{3*} is constant to the left of $-\beta$. Thus, from (9.14)-(9.16) we obtain

$$\begin{aligned} f_{3*}^2(\alpha) = f_{3*}^2(\xi) = f_{3*}(-\beta) &= \mu(1+\sin\theta) \\ &\geqq \mu(1-\sin\theta) = -f_{3*}(\xi) \geqq \beta \geqq \alpha. \end{aligned}$$

The required inequality $f_{3*}^2(\alpha) \geqq \alpha$, from which the implication (9.13) follows, is now proved.

We now calculate the quantity

$$(f_{3*}^2)(\xi) = (f_{3*}^2)(\alpha) = f_{3*}(-\mu(1-\sin\theta)) \tag{9.17}$$

and compare it with either α or ξ, as described above. Let $h_3(\mu, \theta)$ denote $(f_{3*}^2)(\alpha)$; then we easily see

$$h_3(\mu, \theta) = \begin{cases} \mu[\sin\theta - \sin(\theta - \mu(1-\sin\theta))] & \text{if } \mu \leqq \dfrac{\pi/2+\theta}{1-\sin\theta}, \\ \mu(1+\sin\theta) & \text{if } \mu \geqq \dfrac{\pi/2+\theta}{1-\sin\theta}. \end{cases}$$

A simple calculation shows that $\partial h_3(\mu, \theta)/\partial\mu > 0$ for all $\mu > 0$ in the two ranges of μ for which h_3 is defined. Thus $h_3(\mu, \theta)$ is strictly increasing in μ, for each fixed $\theta \in [0, \pi/2)$, and assumes every positive value exactly once for $\mu > 0$. Thus there exist continuous functions $\rho_3, \tau_3 : [0, \pi/2) \to (0, \infty)$ satisfying

$$h_3(\rho_3(\theta), \theta) = \alpha = \pi - \theta, \qquad h_3(\tau_3(\theta), \theta) = \xi = \pi/2 - \theta.$$

By Theorem 7.2 we also have

$$h_3(1/\cos, \theta) < \xi,$$

since $f_3'(0) = -1$ when $\mu = 1/\cos\theta$. Thus,

$$1/\cos\theta < \rho_3(\theta) < \tau_3(\theta)$$

and the result follows immediately. □

As has already been noted in § 3, if the function f_3 is not in our normal form, there may be some difficulties in determining the ranges of the original parameters for which (I) or (III) is satisfied. To illustrate this point, we consider

$$\varepsilon\dot{x}(t) = -x(t) + \mu(1-\sin(x(t-1))), \tag{9.18}$$

which has been studied numerically by Chow and Green [4]. For each $\mu > 0$, we can easily see that

$$\mu(1-\sin x) = x, \qquad 0 < x < \pi/2, \tag{9.19}$$

has a unique fixed point $\theta = \theta(\mu) \in (0, \pi/2)$, and using the implicit function theorem we can see that $\theta'(\mu) > 0$ for $\mu > 0$. The question is does $\mu(1-\sin x)$ satisfy condition (I) or (III) at $x = \theta(\mu)$.

PROPOSITION 9.1. *For each $\mu > 0$, the function $\mu f(x) = \mu(1-\sin x)$ has a unique fixed point $\theta = \theta(\mu) \in (0, \pi/2)$. There exist numbers μ_0 (μ_0 is approximately equal to* 1.1773) *and μ_1 (μ_1 is approximately equal to* 2.3879) *such that $\mu f(x)$ satisfies* (I) *at*

$\theta(\mu)$ if and only if $\mu_0 < \mu < \mu_1$ and $\mu f(x)$ satisfies (III) *at $\theta(\mu)$ if and only if $\mu_0 < \mu \leqq \pi/2$. The equation*

$$\frac{\theta \cos \theta}{1-\sin \theta} = 1, \qquad 0 < \theta < \frac{\pi}{2},$$

has a unique solution $\theta_0 \in (0, \pi/2)$ and $\mu_0 = \theta_0/(1-\sin \theta_0)$. The equation

$$\theta + \left(\frac{\theta}{1-\sin \theta}\right) = \pi, \qquad 0 < \theta < \frac{\pi}{2},$$

has a unique solution $\theta_1 \in (0, \pi/2)$ and $\mu_1 = \theta_1/(1-\sin \theta_1)$.

Proof. The idea of the proof is to parameterize by the fixed point $\theta \in (0, \pi/2)$ instead of by μ. If $\theta \in (0, \pi/2)$ is the fixed point of $\mu f(x)$, then

$$\mu = \frac{\theta}{1-\sin \theta}.$$

Thus, for $0 < \theta < \pi/2$, define $g(x, \theta) = g_\theta(x)$ by

$$g(x, \theta) = g_\theta(x) = \left(\frac{\theta}{1-\sin \theta}\right)(1 - \sin(x+\theta)) - \theta.$$

We can easily check that the map $\theta \to (\theta/1-\sin \theta))$ is strictly increasing for $0 < \theta < \pi/2$ and

$$\{\mu : \mu f(x) \text{ satisfies (III) at } \theta(\mu)\}$$
$$= \{\theta/(1-\sin \theta) : 0 < \theta < \pi/2 \text{ and } g_\theta(x) \text{ satisfies (III) at zero}\},$$

with an analogous equation concerning (I). If the numbers $\alpha = \alpha(\theta), \beta = \beta(\theta), \xi = \xi(\theta)$, and $\eta = \eta(\theta)$ are as in the definition of condition (PM) for g_θ, then $\alpha = \pi - 2\theta$, $\beta = \pi + 2\theta$, $\xi = \pi/2 - \theta$, and $\eta = \pi/2 + \theta$. The same argument as in Theorem 9.1 shows that g_θ satisfies (III) if and only if

$$(g_{\theta^*}^2)\left(\frac{\pi}{2} - \theta\right) \leqq \frac{\pi}{2} - \theta, \tag{9.20}$$

$$g_\theta'(0) = -\left(\frac{\theta \cos \theta}{1-\sin \theta}\right) < -1. \tag{9.21}$$

Since $g_{\theta^*}(\pi/2 - \theta) = -\theta > -\eta$, we easily compute that (9.20) holds if and only if

$$\left(\frac{\theta}{1-\sin \theta}\right) \leqq \frac{\pi}{2}. \tag{9.22}$$

Thus $g_\theta, 0 < \theta < \pi/2$, satisfies (III) if and only if (9.21) and (9.22) hold. Similarly, we see that g_θ satisfies (I) if and only if (9.21) is valid and

$$(g_{\theta^*}^2)(\pi - 2\theta) = \left(\frac{\theta}{1-\sin \theta}\right) - \theta < \pi - 2\theta, \quad \text{that is,}$$
$$\left(\frac{\theta}{1-\sin \theta}\right) + \theta < \pi, \qquad 0 < \theta < \tfrac{\pi}{2}. \tag{9.23}$$

It remains to study where (9.21)-(9.23) hold. Obviously the function $(\theta/(1-\sin \theta)) + \theta = h(\theta)$ satisfies $h(0) = 0$, $\lim_{\theta \to \pi/2} h(\theta) = \infty$ and $h'(\theta) > 0$ for $0 < \theta < \pi/2$, so there is a unique number θ_1 such that

$h(\theta_1) = \pi$ and $h(\theta) < \pi$ if and only if $0 < \theta_1 < \pi$.

In order to study where $k(\theta)=(\theta\cos\theta/(1-\sin\theta))>1$ for $\theta\in(0,\pi/2)$, first observe that the mean value theorem gives

$$1-\sin\theta=\left(\frac{\pi}{2}-\theta\right)\cos(\psi),\qquad \theta<\psi<\frac{\pi}{2}.$$

If $\pi/4\leqq\theta\leqq\pi/2$, this implies

$$k(\theta)=\left(\frac{\theta}{\pi/2-\theta}\right)\left(\frac{\cos\theta}{\cos\psi}\right)>1.$$

On the other hand, if $0<\theta<\pi/4$ we have

$$k'(\theta)=[(\cos\theta-\theta\sin\theta)(1-\sin\theta)+\theta\cos^2\theta](1-\sin\theta)^{-2}.$$

Because we also have

$$\cos\theta-\theta\sin\theta>\cos\theta-\sin\theta\geqq 0\quad\text{for } 0<\theta\leqq\pi/4,$$

we conclude that $k'(\theta)>0$ for $0<\theta\leqq\pi/4$. From the above facts it follows that $k(\theta)=1$ has a unique solution $\theta_0\in(0,\pi/2)$, that $0<\theta_0<\pi/4$, and that $k(\theta)>1$ for $\theta\in(0,\pi/2)$ if and only if $\theta_0<\theta<\pi/2$. This completes the proof of the proposition. (Approximate values of θ_0 and θ_1, and hence of μ_0 and μ_1, can easily be computed using Newton's method.) □

Proposition 9.1 and the results summarized in § 2 provide some explanation of the numerical results in [4]. For example, if $\mu_0<\mu\leqq\pi/2$ we see for small ε the regular SOP-solutions predicted by Theorem 2.3. Theorem 2.1 asserts that SOP-solutions persist for $\mu_0<\mu<\mu_1$ and $\varepsilon>0$ sufficiently small. However, these solutions apparently lose stability for μ near μ_1 and $\varepsilon>0$ small: for μ near μ_1 Chow and Green appear to have found, numerically, periodic solutions that are *not* SOP-solutions.

We now want to examine when (I), (II), or (III) is satisfied by the functions f_4 or f_5. With $k=4$ or $k=5$ and fixed parameters $\nu\geqq 0$ and $\lambda>\nu+1$ (when $k=5$), we will write

$$f_k(x)=\mu\bar{f}_k(x), \tag{9.24}$$

where the function $\bar{f}_k(x)$ does not depend on μ. In our next several theorems we will discuss the range of μ for which the functions f_4 and f_5 satisfy (I) or (III), and we will return later to (II). The next theorem provides a reasonably sharp and general answer concerning when (III) holds for functions $\mu\bar{f}(x)$; the question of (I) seems more difficult.

THEOREM 9.2. *Assume $\bar{f}:(0,\infty)\to(0,\infty)$ is a C^3 function and there exists a number $\theta\geqq 0$ such that $\bar{f}'(s)>0$ for $0<s<\theta$ and $\bar{f}'(s)<0$ for $s>\theta$. Assume there exists a unique number $s_0>\theta$ such that*

$$\frac{d}{ds}(s\bar{f}(s))|_{s=s_0}=0,$$

and assume also

$$\frac{d}{ds}(s\bar{f}(s))|_{s=s_0}<0\quad\forall s>s_0.$$

Finally, assume $(S\bar{f})(x)<0$ for $x>\theta$. Define functions $\mu(s)$, $x_1(s)$, $x_2(s)$, and $x_3(s)$ by $\mu(s)=s(\bar{f}(s))^{-1}$, $x_1(s)=\mu(s)\bar{f}(\theta)$, and

$$x_j(s)=\mu(s)\bar{f}(x_{j-1}(s))\quad\text{for } j\geqq 2.$$

If $x_0(\mu)$ denotes the unique fixed x of $\mu f(x)$ such that $x \geqq \theta$ (for $\mu \geqq \theta(\bar{f}(\theta))^{-1}$), then

$$\{\mu\colon \mu\bar{f}(x) \text{ satisfies (I) at } x_0(\mu)\} = \{\mu(s)\colon s > s_0 \text{ and } x_3(s) > s\}$$

and

$$\{\mu\colon \mu f(x) \text{ satisfies (III) at } x_0(\mu)\} = \{\mu(s)\colon s > s_0 \text{ and } x_2(s) \geqq \theta\}.$$

There exists a number $\rho > \mu(s_0) \equiv \mu_0$ such that

$$\{\mu\colon \mu f(x) \text{ satisfies (III) at } x_0(\mu)\} = (\mu_0, \rho],$$

and $\rho < \infty$ if $\lim_{x\to\infty} x\bar{f}(x) = 0$. The number ρ is $z(\bar{f}(\theta))^{-1}$ where z is the unique solution of $z\bar{f}(z) = \theta\bar{f}(\theta)$ such that $z > s_0$ (if such a solution exists).

If D is an open subset of $\mathbb{R}^m$ and $\bar{f}\colon (0,\infty)\times D \to (0,\infty)$ is a C^3 map such that $x \to \bar{f}(x,\gamma)$ satisfies the conditions of our theorem for each $\gamma \in D$ (so $\theta = \theta(\gamma)$ and $s_0 = s_0(\gamma)$ exist and are easily proven to be continuous), then the number $\rho = \rho(\gamma)$ is also a continuous function of γ. (If $\rho(\gamma) = \infty$ for some γ, continuity is interpreted in the obvious way.)

Proof. First assume $\bar{f}$ is independent of $\gamma \in D$. For a given $\mu > 0$ let $x_0 = x_0(\mu) \geqq \theta$ and define $\delta = \delta(\mu) \leqq \theta$ by

$$\bar{f}(\delta) = \bar{f}(x_0).$$

Theorem 4.1 implies that $\mu\bar{f}(x)$ satisfies condition (0) at x_0 if and only if $\mu > \mu_0 = s_0(\bar{f}(s_0))^{-1} = \mu(s_0)$, and the results of § 7 imply that $\mu\bar{f}(x)$ satisfies (III) at x_0 if and only if $\mu > \mu_0$ and

$$(\mu\bar{f})^2(\theta) \geqq \theta \tag{9.25}$$

where $(\mu\bar{f})^j$ is the composition of $\mu\bar{f}$ with itself j times. Similarly, $\mu\bar{f}(x)$ satisfies (I) at x_0 if and only if $\mu > \mu_0$ and

$$(\mu\bar{f})^2(\theta) > \delta = \delta(\mu),$$

or equivalently $\mu > \mu_0$ and

$$(\mu\bar{f})^3(\theta) > x_0(\mu). \tag{9.26}$$

(Note that $\mu\bar{f}(\theta) > \mu\bar{f}(x_0) = x_0$, so $(\mu\bar{f})^2(\theta) < x_0$.)

As in Theorem 4.1 the key idea is to parameterize by $s = x_0(\mu)$ instead of μ. Since $\mu\bar{f}(x_0) = x_0$, this gives

$$\mu = s(\bar{f}(s))^{-1} = \mu(s),$$

which is an increasing function of s for $s \geqq \theta$. In terms of this parameterization we find that $\mu(s)\bar{f}(x)$ satisfies (III) at $s > \theta$ if and only if $s > s_0$ and

$$x_2(s) \geqq \theta. \tag{9.27}$$

Similarly, we see that $\mu(s)\bar{f}(x)$ satisfies (I) at s if and only if $s > s_0$ and

$$x_3(s) > s. \tag{9.28}$$

Since Theorem 7.2 and Corollary 7.1 imply that there exists $\delta > 0$ such that $\mu(s)\bar{f}(x)$ satisfies (III) for $s_0 < s \leqq s_0 + \delta$, we have proved the first part of the theorem.

It remains to prove that (III) is satisfied on an interval $(\mu_0, \rho]$. Equivalently, it suffices to prove that

$$\{s\colon s > s_0, x_2(s) \geqq \theta\}$$

is an interval. We know that $x_2(s) > \theta$ for s near s_0, so it suffices to prove $x_2'(s) < 0$ for $s > s_0$. A calculation gives

$$x_1'(s) = \bar{f}(\theta)(\bar{f}(s))^{-2}[\bar{f}(s) - s\bar{f}'(s)],$$

and since $\bar{f}'(s) < 0$ for $s > s_0$, $x_1'(s) > 0$ for all $s > s_0$. Note that we have

$$\mu(s)\bar{f}(\theta) = x_1(s) > \mu(s)\bar{f}(s) = s > s_0,$$

so we find that

$$\lim_{s\to\infty} x_1(s) = \infty.$$

We can write

$$\mu(s) = x_1(s)(\bar{f}(\theta))^{-1},$$

so if we define $g(x) = (\bar{f}(\theta))^{-1}x\bar{f}(x)$,

$$x_2(s) = g(x_1(s)), \qquad x_2'(s) = g'(x_1(s))x_1'(s).$$

However, $x_1(s) > s_0$ and we have assumed that $g'(x) < 0$ for $x > s_0$, so $x_2'(s) < 0$.

If $\sigma > s_0$ is such that $x_2(\sigma) = \theta$, and if $z = x_1(\sigma)$ the above calculation shows that $z\bar{f}(z) = \theta\bar{f}(\theta)$ and $\rho = \mu(\sigma) = z(\bar{f}(\theta))^{-1}$.

If we assume $\lim_{x\to\infty} x\bar{f}(x) = 0$, we obtain (because $x_1(s) \to \infty$ as $s \to \infty$) the following

$$\lim_{s\to\infty} x_2(s) = \lim_{s\to\infty} g(x_1(s)) = 0.$$

This implies that for all large s, $x_2(s) < \theta$, so ρ is finite in this case.

If $\bar{f}$ depends on a parameter $\gamma \in D$, the continuity of $\rho(\gamma)$ follows easily from the implicit function theorem at points where $\rho(\gamma) < \infty$, and continuity at points γ where $\rho(\gamma) = \infty$ is also easy. Details are left to the reader. □

Remark (9.1). *If* $\bar{f}: (0, \infty) \to (0, \infty)$ is as Theorem 9.2 and $\varphi: (\alpha, \beta) \to (0, \infty)$ is a C^1 map onto $(0, \infty)$ with positive derivative, the remarks at the beginning of this section show that $\varphi^{-1}(\mu\bar{f})\varphi$ satisfies (I) or (III) if and only if $\mu\bar{f}$ satisfies (I) or (III).

As an immediate consequence of Theorem 9.2 we obtain Theorem 9.3.

THEOREM 9.3. *Let* $f_k(x)$, $k = 4$ *or* 5, *be as usual and assume* $\nu > 0$ *and* $\lambda > \nu + 1$ *if* $k = 5$. *Define* $\mu_k(s) = s(\bar{f}_k(s))^{-1}$ *and define* $\theta_k = \nu$ *for* $k = 4$ *and* $\theta_k^\lambda = \nu(\lambda - \nu)^{-1}$ *for* $k = 5$. *Define* $\sigma_k = \nu + 1$ *for* $k = 4$ *and* $\sigma_k^\lambda = (\nu + 1)(\lambda - v - 1)^{-1}$ *for* $k = 5$. *Define* z_k *to be the unique solution* $z > \sigma_k$ *of*

$$z\bar{f}_k(z) = \theta_k\bar{f}_k(\theta_k).$$

If x_0 *denotes generically the unique fixed point greater than* θ_k *of* $\mu\bar{f}_k(x)$, *then*

$$\{\mu: \mu\bar{f}_k(x) \text{ satisfies (III) at } x_0\} = (\mu_k(\sigma_k), \rho_k],$$

where $\rho_k = z_k(\bar{f}_k(\theta_k))^{-1}$ *is a continuous function of* ν *and* λ. *If* $x_1(s) = \mu_k(s)\bar{f}_k(\theta_k)$ *and* $x_j(s) = \mu_k(s)\bar{f}_k(x_{j-1}(s))$ *for* $j \geqq 2$, *then*

$$\{\mu: \mu\bar{f}_k(x) \text{ satisfies (I) at } x_0\} = \{\mu_k(s): s > \sigma_k \text{ and } x_3(s) > s\}.$$

If $\nu = 0$, *then* $\mu\bar{f}_k(x)$ *satisfies* (III) *at* x_0 *for all* $\mu > \mu_0$.

Proof. The number θ_k plays the role of θ in Theorem 9.2, and σ_k the role of s_0. We have already verified the negative Schwarzian condition on $\bar{f}_k$ and the other hypotheses of Theorem 9.2 are easily verified, so Theorem 9.3 follows directly from Theorem 9.2. □

We now want to study more precisely when $\mu\bar{f}_4(x)$ satisfies (I). It is convenient to give a calculus lemma first.

LEMMA 9.1. *If* $\nu > 0$, *then*

$$\left(\frac{\nu+2}{\nu}\right)^{\nu+1} > e^2, \tag{9.29}$$

and if $\nu \geqq 1$

$$e < \left(\frac{\nu+2}{\nu+1}\right)\left(\frac{\nu+1}{\nu}\right)^{\nu}. \tag{9.30}$$

If $0 \leqq c < \sqrt{5}-2$, *then there exists* $\nu(c) \geqq 1$ *such that the following inequality is valid for* $\nu \geqq \nu(c)$:

$$e^{1+c} \leqq \left(\frac{\nu+1+c}{\nu}\right)^{\nu}\left(\frac{\nu+2}{\nu+1+c}\right). \tag{9.31}$$

Proof. By taking natural logarithms we see that (9.29) is equivalent to proving

$$(\nu+1)\log\left(1+\frac{2}{\nu}\right) > 2,$$

and the above inequality is equivalent to

$$(\nu+1)\int_0^{2/\nu}\left(\frac{1}{1+t}\right)dt > \int_0^{2/\nu} \nu\, dt.$$

The above inequality is equivalent to

$$\int_0^{1/\nu}\left(\frac{1-\nu t}{1+t}\right)dt = I_1 > \int_{1/\nu}^{2/\nu}\left(\frac{\nu t-1}{1+t}\right)dt = I_2.$$

Change variables in I_1 by setting $t = 1/\nu - \rho$ to obtain

$$I_1 = \int_0^{1/\nu}\left(\frac{\nu\rho}{1+(1/\nu)-\rho}\right)d\rho$$

and change variables in I_2 by setting $t = 1/\nu + \rho$ to obtain

$$I_2 = \int_0^{1/\nu}\left(\frac{\nu\rho}{1+(1/\nu)+\rho}\right)d\rho.$$

Since $\nu > 0$, we have

$$\nu\rho(1+\nu^{-1}-\rho)^{-1} > \nu\rho(1+\nu^{-1}+\rho)^{-1} \quad \text{for } 0 < \rho < \nu^{-1},$$

so $I_1 > I_2$.

The proofs of (9.30) and (9.31) are like the proof of (9.29). If $\nu \geqq 1$, (9.30) is equivalent to

$$1 < \log\left(1+\frac{1}{\nu+1}\right) + \nu\log\left(1+\frac{1}{\nu}\right).$$

Expressing both sides as integrals, the above inequality is equivalent to proving

$$\int_0^{1/\nu} \nu\, dt < \int_0^{1/\nu}\left(\frac{\nu+1}{1+t}\right)dt - \int_{1/(\nu+1)}^{1/\nu}\left(\frac{1}{1+t}\right)dt.$$

Simplification shows that the above inequality is equivalent to

$$0 < \int_0^{1/(\nu+1)}\left(\frac{1-\nu t}{1+t}\right)dt - \int_{1/(\nu+1)}^{1/\nu}\left(\frac{\nu t}{1+t}\right)dt = J_1 - J_2. \tag{9.32}$$

Making the change of variables $t = 1/\nu - \rho$ in J_2, we have

$$J_2 = \int_0^{1/\nu - 1/(\nu+1)} \left(\frac{1-\nu t}{1+\nu^{-1}-t}\right) dt. \tag{9.33}$$

Since $\nu \geqq 1$ we can easily verify that

$$(\nu+1)^{-1} \geqq \nu^{-1} - (\nu+1)^{-1},$$

and using (9.32) and (9.33) we see that $J_1 > J_2$ if

$$(1-\nu t)(1+t)^{-1} > (1-\nu t)(1+\nu^{-1}-t)^{-1} \quad \text{for } 0 < t < \nu^{-1} - (\nu+1)^{-1}. \tag{9.34}$$

However, again using that $\nu \geqq 1$, we can check that inequality (9.34) holds, so $J_1 > J_2$.

By taking logarithms we see that inequality (9.31) is equivalent to

$$1 + c \leqq \nu \log\left(1 + \frac{1+c}{\nu}\right) + \log\left(1 + \frac{1-c}{\nu+1+c}\right),$$

or, by expressing both sides as integrals,

$$\int_0^{(1+c)/\nu} \nu \, dt \leqq \int_0^{(1+c)/\nu} \left(\frac{\nu}{1+t}\right) dt + \int_0^{(1-c)/(\nu+1+c)} \left(\frac{1}{1+t}\right) dt.$$

By simplifying we see that inequality (9.31) is equivalent to

$$K_1 = \int_0^{(1+c)/\nu} \left(\frac{\nu t}{1+t}\right) dt \leqq \int_0^{(1-c)/(\nu+1+c)} \left(\frac{1}{1+t}\right) dt = K_2.$$

Since $(1+t)^{-1} < 1$ for $t > 0$ we see that

$$K_1 < \int_0^{(1+c)/\nu} (\nu t)\, dt = \left(\frac{1}{2}\right)\frac{(1+c)^2}{\nu}. \tag{9.35}$$

On the other hand, $(1+t)^{-1} > 1 - t$, so

$$K_2 > \int_0^{(1-c)/(\nu+1+c)} (1-t)\, dt = \left(\frac{1-c}{\nu+1+c}\right) - \frac{1}{2}\left(\frac{1-c}{\nu+1+c}\right)^2. \tag{9.36}$$

For a given c, $0 < c < 1$, it follows from (9.35) and (9.36) that $K_1 < K_2$ if

$$\left(\frac{1}{2}\right)\frac{(1+c)^2}{\nu} < \left(\frac{1-c}{\nu+1+c}\right) - \frac{1}{2}\left(\frac{1-c}{\nu+1+c}\right)^2.$$

Multiplying by $2\nu(\nu+1+c)$, we see that the above inequality is equivalent to

$$0 < (1-4c-c^2)\nu - (1+c)^3 - (1-c)^2\left(\frac{\nu}{\nu+1+c}\right). \tag{9.37}$$

If $1-4c-c^2 > 0$, i.e., $c < \sqrt{5}-2$, then inequality (9.37) will be satisfied for all $\nu \geqq \nu(c)$, where

$$\nu(c) = (1-4c-c^2)^{-1}[(1+c)^3 + (1-c)^2] = (1-4c-c^2)^{-1}[c^3+4c^2+c+2]. \tag{9.38}$$ □

THEOREM 9.4. *There exists a continuous function*

$$\tau_4 : (0, \infty) \to (0, \infty)$$

satisfying $\mu_0(\nu) = (\nu+1)^{1-\nu} e^{\nu+1} < \rho_4(\nu) < \tau_4(\nu)$, *where* $\rho_4 = \rho_4(\nu)$ *is as in Theorem* 9.3, *such that* (I) *holds for* $\mu \bar{f}_4(x)$ (*at its unique fixed point in* (ν, ∞)) *if and only if* $\mu_0(\nu) < \mu < \tau_4(\nu)$. *Furthermore, we have that*

$$\lim_{\nu \to 0^+} \tau_4(\nu) = \infty \text{ and } \lim_{\nu \to \infty} [\tau_4(\nu) - \mu_0(\nu)] = 0, \text{ and for } \nu \geqq 2 \tag{9.39}$$

$$\tau_4(\nu) - \mu_0(\nu) \leqq \mu(2\nu) - \mu(\nu+1) \leqq (e^2/2\nu)^\nu (2\nu),$$

where we have defined $\mu(s) = s^{1-\nu} e^s$. *For* $0 < c < \sqrt{5} - 2$ *and for* $\nu \geqq \nu(c)$ ($\nu(c)$ *as in* (9.38)) *we have*

$$[\mu_0(\nu), \rho_4(\nu)] \supset \{s^{1-\nu} e^s : \nu + 1 < s \leqq \nu + 1 + c\}. \tag{9.40}$$

Proof. For notational convenience we write $\bar{f}_4(x) = \bar{f}(x)$, $\theta = \nu$, and $s_0 = \nu + 1$, and define $x_j(s)$ as in Theorem 9.2. Theorems 9.2 and 9.3 imply

$$[\mu_0(\nu), \rho_4(\nu)] = \{\mu(s) : x_2(s) \geqq \theta\},$$

$$\{\mu : \mu \bar{f}(x) \text{ satisfies (I)}\} = \{\mu(s) : s > s_0 \text{ and } x_3(s) > s\}.$$

Furthermore, we have already shown that

$$x_1'(s) > 0, \quad x_2'(s) < 0 \quad \forall s > s_0.$$

We fix $\nu > 0$ and first show $x_3(s) < s$ for all large s or, equivalently, that

$$\log (s^{-1} x_3(s)) < 0 \quad \forall \text{ large } s.$$

Using the definition of $x_j(s)$, we find

$$\begin{aligned} \log (s^{-1} x_3(s)) &< -\nu \log s + s + \nu \log x_2 \\ &= -\nu^3 \log s + (1 + \nu + \nu^2) s + \nu^3 \log \nu - \nu^3 - \nu^{\nu+1} e^{-\nu} s^{1-\nu} e^s. \end{aligned} \tag{9.41}$$

Because the e^s term is dominant for large s, the right-hand side of (9.41) is negative for all large s, so (for fixed $\nu > 0$) for every sufficiently large μ, $\mu \bar{f}(x)$ does not satisfy (I).

If $\nu \geqq 1$, it is a calculus exercise (which we leave to the reader) to prove that the derivative of the right-hand side of (9.41) with respect to s is negative for $s \geqq 2\nu$. Another calculus exercise left to the reader is to verify that

$$\frac{d}{d\nu} (\log (s^{-1} x_3(s))|_{s=2\nu}) < 0 \quad \text{for } \nu \geqq 2.$$

A direct calculation shows that the right-hand side of (9.41) is negative for $\nu = 2$ and $s = 4 = 2\nu$, and combining the above information we conclude that

$$\log (s^{-1} x_3(s)) < 0 \quad \text{for } s \geqq 2\nu \text{ and } \nu \geqq 2. \tag{9.42}$$

It follows from inequality (9.42) that, for $\nu \geqq 2$,

$$\{\mu : \mu \bar{f}(x) \text{ satisfies (I)}\} \subset \{\mu(s) : \nu + 1 < s < 2\nu\} = (\mu(\nu + 1), \mu(2\nu)).$$

Because $\mu(s)$ is increasing for $s > \nu - 1$ we have

$$\begin{aligned} \mu(2\nu) - \mu(\nu + 1) < \mu(2\nu) - \mu(\nu) &= \nu \left(\frac{e}{\nu}\right)^{\nu} \left[2\left(\frac{e}{2}\right)^{\nu} - 1\right] \\ &\leqq (2\nu)\left(\frac{e^2}{2\nu}\right)^{\nu}, \end{aligned}$$

which immediately gives (9.39) (although we have not yet proved that $\tau_4(\nu)$ exists).

To prove that $\tau_4(\nu) \to \infty$ as $\nu \to 0^+$ (assuming the existence of $\tau_4(\nu)$), it suffices

to prove that given any $M>0$, there exists $\delta=\delta(M)$ such that for $0<\nu<\delta(M)$ and $\nu+1<s\leqq M$,

$$\log(s^{-1}x_3(s))=-\nu^3\log s+(1+\nu+\nu^2)s+\nu^3\log\nu-\nu^3-\nu^{\nu+1}s^{1-\nu}e^{s-\nu}-x_2>0. \tag{9.43}$$

Because $\nu^\nu e^{-\nu}$ converges to one as $\nu\to 0^+$, we see that $x_1(s)=s^{1-\nu}e^s\nu^\nu e^{-\nu}$ converges uniformly on $[1, M]$ to se^s as $\nu\to 0^+$, and using this we see that

$$\lim_{\nu\to 0^+}x_2(s)=\lim_{\nu\to 0^+}s^{1-\nu}e^s x_1^\nu e^{-x_1}=s\exp(s-se^s),$$

and that the convergence is uniform in $s\in[1, M]$. Using this information, we see that

$$\lim_{\nu\to 0^+}(\log(s^{-1}x_3(s)))=s-s\exp(s-se^s), \tag{9.44}$$

and that the convergence is uniform in $s\in[1, M]$. Since the right-hand side of (9.44) is positive on $(0,\infty)$, there exists $\delta=\delta(M)>0$ such that

$$\log(s^{-1}x_3(s))>0 \quad \text{for } 1\leqq s\leqq M,\quad 0<\nu<\delta.$$

We need only prove the existence of $\tau_4(\nu)$, or equivalently that

$$\{s>\nu+1:\log(s^{-1}x_3(s))<0\} \quad \text{is an interval.}$$

It is convenient to make the following observation first.

We claim that if $x_1(s)\leqq\nu+2$ and $s>\nu+1$, then $x_2(s)>\nu$. Because $x_1'(s)>0$ and $x_2'(s)<0$ for $s>\nu+1$, it suffices to prove that if $x_1(s)=\nu+2$, then $x_2(s)>\nu$. However, if $x_1(s)=\nu+2$, we find as in the proof of Theorem 9.2, that

$$x_2=x_1\bar f(x_1)(\bar f(\theta))^{-1}=\nu^{-\nu}e^\nu(\nu+2)(\nu+2)^\nu e^{-(\nu+2)},$$

so $x_2(s)>\nu$ if

$$\left(\frac{\nu+2}{\nu}\right)^{\nu+1}>e^2,$$

which is (9.25). We conclude that if $x_1(s)\leqq\nu+2$, then $\mu(t)\bar f(x)$ satisfies (III) for $\nu+1<t\leqq\nu+2$. A calculation shows that $x_1(\nu+1+c)\leqq\nu+2$ if and only if

$$e^{1+c}\leqq\left(\frac{\nu+1+c}{\nu}\right)^\nu\left(\frac{\nu+2}{\nu+1+c}\right), \tag{9.45}$$

and Lemma 9.1 implies that if $0<c<\sqrt{5}-2$ and $\nu\geqq\nu(c)$, inequality (9.45) holds. This proves the inclusion (9.40).

Logarithmic differentiation easily yields the following formulas:

$$\begin{aligned}
&\frac{dx_1}{ds}=x_1\left[\frac{s+1-\nu}{s}\right],\\
&\frac{dx_2}{ds}=x_2\left[\frac{s+1-\nu}{s}\right][\nu+1-x_1],\\
&\frac{d}{ds}\left(\log\left(\frac{x_3(s)}{s}\right)\right)=\left(\frac{s+1-\nu}{s}\right)[1+(\nu-x_2)(\nu+1-x_1)]-\left(\frac{1}{s}\right).
\end{aligned} \tag{9.46}$$

Define s_* to be the first $s>\nu+1$ such that $\log(s^{-1}x_3(s))=0$, so we know

$$\frac{d}{ds}\log(s^{-1}x_3(s))|_{s=s_*}=s_*^{-1}[(s_*+1-\nu)(1-(\nu-x_2)(x_1-\nu-1))-1]\leqq 0. \tag{9.47}$$

Theorem 9.3 implies $x_2(s_*) < \nu$, and the remarks above show $x_1(s_*) > \nu + 2$. To complete the proof we need only show

$$(s+1-\nu)[1-(\nu-x_2(s))(x_1(s)-\nu-1)]-1 \equiv \phi(s) < 0$$

for $s > s_*$, and because $\phi(s_*) \leqq 0$, it suffices to prove

$$\begin{aligned} \phi'(s) &= (\nu-x_2)(x_1-\nu-1)-(s+1-\nu)^2 s^{-1}(x_1-\nu-1)^2 x_2 \\ &\quad -(s+1-\nu)^2 x^{-1} x_1(\nu-x_2) < 0 \end{aligned} \tag{9.48}$$

for $s > s_*$.

Case 1. Assume $\nu \geqq 1$. Using the estimates $x_1 - \nu - 1 > 1$ and $\nu - x_2 > 0$ in the formula for $\phi'(s)$ for $s > s_*$, we obtain

$$\begin{aligned} \phi'(s) &< 1-(\nu-x_2)-x_2(s+1-\nu)^2 s^{-1}-(s+1-\nu)^2 s^{-1}(\nu+2)(\nu-x_2) \\ &= 1-(\nu-x_2)[1+(\nu+1)(s+1-\nu)^2 s^{-1}]-(s+1-\nu)^2 s^{-1}\nu. \end{aligned}$$

The previous inequality shows

$$\phi'(s) < 1-(s+1-\nu)^2 s^{-1}\nu. \tag{9.49}$$

The function on the right-hand side of (9.49) is decreasing for $s > \nu+1$, so inequality (9.49) implies that, for $s \geqq s_*$,

$$\phi'(s) < 1-\frac{4\nu}{\nu+1} < 0.$$

Case 2. Assume $0 < \nu < 1$. Because x_2 is decreasing and less than ν for $s \geqq s_*$ and x_1 is increasing and greater than $\nu+1$ for $s > \nu+1$, $(\nu - x_2)(x_1 - \nu - 1)$ is an increasing function of s for $s \geqq s_*$. At $s = s_*$, inequality (9.47) implies

$$\begin{aligned} &\tfrac{1}{2} < 1-(s_*+1-\nu)^{-1} \leqq (\nu-x_2(s_*))(x_1(s_*)-\nu-1), \quad \text{so} \\ &\tfrac{1}{2} < (\nu-x_2(s))(x_1(s)-\nu-1) \quad \text{for } s \geqq s_*. \end{aligned} \tag{9.50}$$

Using the equation for $\phi'(s)$ in (9.48), we see

$$\phi'(s) < 1-(\nu-x_2)(x_1-1-\nu)-(s+1-\nu)^2 s^{-1}(\nu-x_2)x_1. \tag{9.51}$$

Because $0 < \nu < 1$, we have $(s+1-\nu)^2 s^{-1} > s > 1$, so from (9.50) and (9.51) we derive

$$\phi'(s) < 1-(\nu-x_2)(x_1-1-\nu)-(\nu-x_2)(x_1-1-\nu) < 0$$

for $s > s_*$. The proof is now complete. □

Next we want to analyze when $\mu \bar{f}_5(x)$ satisfies (I). Unfortunately, our results are incomplete. We conjecture that there exists a continuous function $\tau_5(\nu, \lambda)$ (allowing $\tau_5(\nu, \lambda) = \infty$) defined for $\nu > 0$ and $\lambda > \nu + 1$ such that $\mu \bar{f}_5(x)$ satisfies (I) if and only if $\mu_0(\nu, \lambda) < \mu < \tau_5(\nu, \lambda)$ (where $\mu_0(\nu, \lambda)$ is as in Table 1). By using Theorem 9.3 and Theorem 9.5, we have given a computer-assisted proof of this conjecture for various specific ν and λ, but we have not proved it in general.

THEOREM 9.5. *Assume $\nu > 0$ and $\lambda > \nu + 1$, and let $x_0 = x_0(\mu, \nu, \lambda)$ and $\mu_0 = \mu_0(\nu, \lambda)$ be as in Table 1 for the function $f_5(x) = \mu \bar{f}_5(x)$. The function $\mu \bar{f}_5(x)$ satisfies* (I) *at x_0 for all large μ if*

$$\nu + 1 < \lambda \leqq \nu + \left(\frac{1}{2\nu}\right) + \left(\frac{1}{2\nu}\right)\sqrt{4\nu^2+1} \equiv \phi(\nu), \tag{9.52}$$

while if $\lambda > \phi(\nu)$, there exists a number $\gamma = \gamma(\nu, \lambda) < \infty$ such that $\mu \bar{f}_5(x)$ does not satisfy (I) *at x_0 for any $\mu > \gamma(\nu, \lambda)$. If $\nu + 1 < \lambda \leqq \nu + 1 + (1/2\nu)$, $\mu \bar{f}_5(x)$ satisfies condition* (I) *at x_0 for all $\mu > \mu_0(\nu, \lambda)$.*

Proof. While not essential, a change of variables will simplify our calculations. For $x>0$, define $\psi(x)=x^{1/\lambda}$, so

$$(\psi^{-1}(\mu\bar{f}_5)\psi)(x)=\mu^\lambda x^\nu(1+x)^{-\lambda}\equiv\mu^\lambda\bar{h}_5(x).$$

The remarks at the beginning of this section show that $\mu\bar{f}_5(x)$ satisfies (I) at x_0 if and only if $\mu^\lambda\bar{h}_5(x)$ satisfies (I) at x_0^λ. As in Theorem 9.2 we see that $\mu^\lambda\bar{h}_5(x)$ satisfies (I) at x_0^λ if and only if $\mu>\mu_0$ and

$$(\mu^\lambda\bar{h}_5)^3(\theta)>x_0^\lambda, \tag{9.53}$$

where $\theta=\nu(\lambda-\nu)^{-1}$, the point where $\bar{h}_5$ achieves its maximum.

As before, it is convenient to parameterize by $s=x_0^\lambda$, the fixed point of $\mu^\lambda\bar{h}_5$. Define a function $\mu(s)$ by

$$\mu(s)^\lambda=s^{1-\nu}(1+s)^\lambda=s(\bar{h}_5(s))^{-1},$$

so s is a fixed point of $\mu(s)^\lambda\bar{h}_5(x)$. Define $x_1(s)=\mu(s)\bar{h}_5(\theta)$ and $x_j(s)=\mu(s)\bar{h}_5(x_{j-1}(s))$ for $j>1$. Just as in Theorem 9.2 we obtain from (9.53) that

$$\{\mu:\ \mu\bar{f}_5(x)\text{ satisfies (I) at }x_0\}=\{\mu(s):\ s>s_0\text{ and }x_3(s)>s_0\}, \tag{9.54}$$

where $s_0=(\nu+1)(\lambda-\nu-1)^{-1}$ as in (9.51). The proof that $x_1'(s)>0$ and $x_2'(s)<0$ for all $s>s_0$ is as in Theorem 9.2 and is left to the reader.

A calculation shows

$$\begin{aligned}&x_1(s)=x_1=s^{\lambda+1-\nu}(1+s^{-1})^\lambda\theta_1,\quad\text{where}\\&\theta_1=\left(\frac{\nu}{\lambda}\right)^\nu\left(\frac{\lambda-\nu}{\lambda}\right)^{\lambda-\nu}<1.\end{aligned} \tag{9.55}$$

A further calculation yields

$$x_2(s)=x_2=s^{1-(\lambda-\nu)^2}\theta_1^{\nu-\lambda}(1+s^{-1})^{\lambda+\nu\lambda-\lambda^2}(1+x_1^{-1})^{-\lambda}. \tag{9.56}$$

Equation (9.55) shows $\lim_{s\to\infty}x_1(s)=\infty$, and (9.56) gives

$$\lim_{s\to\infty}x_2(s)=\lim_{s\to\infty}s^{1-(\lambda-\nu)^2}\theta_1^{\nu-\lambda}=0,$$

because $\lambda>\nu+1$. Substituting (9.56) in the formula for $x_3(s)$, we obtain

$$\begin{aligned}&s^{-1}x_3(s)=s^{\lambda-\nu(\lambda-\nu)^2}\theta_1^{\nu(\nu-\lambda)}(1+s^{-1})^\delta(1+x_1^{-1})^{-\nu\lambda}(1+x_2)^{-\lambda}\quad\text{where}\\&\delta=\lambda+\nu\lambda+\nu^2\lambda-\nu\lambda^2.\end{aligned} \tag{9.57}$$

Equation (9.56) implies

$$\lim_{s\to\infty}s^{-1}x_3(s)=\lim_{s\to\infty}s^{\lambda-\nu(\lambda-\nu)^2}\theta_1^{\nu(\nu-\lambda)}.$$

Because $\lambda-\nu(\lambda-\nu)^2=0$ for $\lambda=\phi(\nu)$, $\lambda-\nu(\lambda-\nu)^2>0$ for $\nu+1<\lambda<\phi(\nu)$, and $\lambda-\nu(\lambda-\nu)^2<0$ for $\lambda>\phi(\nu)$, we obtain

$$\lim_{s\to\infty}s^{-1}x_3(s)=\begin{cases}\infty & \text{for }\nu+1<\lambda<\phi(\nu),\\ \theta_1^{\nu(\nu-\lambda)} & \text{for }\lambda=\phi(\nu),\\ 0 & \text{for }\lambda>\phi(\nu).\end{cases} \tag{9.58}$$

Using (9.58) and (9.54) and recalling that $0<\theta_1<1$, we obtain the first part of the theorem.

It remains to prove the final part of the theorem. We know (from Theorem 7.2 and the remarks at the beginning of this section) that $s^{-1}x_3(s) > 1$ for $s = s_0$. Therefore, to prove that $\mu\bar{f}_5(x)$ satisfies (I) for all $\mu > \mu_0$ (when $\nu+1 < \lambda \leqq \nu+1+(1/2\nu)$) we need only prove that $s^{-1}x_3(s)$ is an increasing function for $s \geqq s_0$. Because $x_1'(s) > 0$ and $x_2'(s) < 0$ for $s \geqq s_0$, it is clear that $(1+x_1^{-1})^{-\nu\lambda}(1+x_2)^{-\lambda}$ is an increasing function of s. Thus, by using (9.57) we see that $s^{-1}x_3(s)$ is increasing if

$$\frac{d}{ds} s^{\lambda-\nu(\lambda-\nu)^2}(1+s^{-1})^{\lambda+\nu\lambda+\nu^2\lambda-\nu\lambda^2} \geqq 0 \quad \text{for } s \geqq s_0. \tag{9.59}$$

By differentiating logarithmically, we see that inequality (9.59) will hold if

$$s^{-1}(s+1)^{-1}[(\lambda-\nu(\lambda-\nu)^2)(s+1)-\lambda(1+\nu+\nu^2-\nu\lambda)] \geqq 0 \quad \text{for } s \geqq s_0. \tag{9.60}$$

Since $s+1 \geqq \lambda(\lambda-\nu-1)^{-1}$ for $s \geqq s_0$ we see that (9.60) will be satisfied if

$$[\lambda-\nu(\lambda-\nu)^2]\left(\frac{\lambda}{\lambda-\nu-1}\right)-\lambda[1+\nu(1+\nu-\lambda)] \geqq 0. \tag{9.61}$$

Recalling that $\lambda > \nu+1$ and simplifying, we see that (9.61) will be satisfied if

$$\lambda \leqq 1+\nu+(1/2\nu),$$

and this completes the proof. □

If $\nu = 1$, Theorem 9.5 ensures that $\mu\bar{f}_5(x)$ satisfies (I) at x_0 for all $\mu > \mu_0$ if $2 < \lambda \leqq 2.5$, while the number $\phi(1)$ equals $(\frac{1}{2})(3+\sqrt{5})$ or approximately 2.618. We can, however, give an ad hoc argument (which we omit) and prove that, for $\nu = 1$ and $2 < \lambda \leqq (\frac{1}{2})(3+\sqrt{5})$, $\mu\bar{f}_5(x)$ satisfies (I) for all $\mu > \mu_0$.

We now want to study when a function $\mu\bar{f}(x)$ satisfies (II) at a fixed point x_0; our particular interest, of course, is $\bar{f} = \bar{f}_4$ or $\bar{f} = \bar{f}_5$. We first make some preliminary calculations concerning local stability of period 2 points of $\mu\bar{f}(x) = f(x)$.

Suppose, for some μ, there exist numbers $0 < x_1 < x_2$ satisfying

$$\mu\bar{f}(x_1) = x_2 \quad \text{and} \quad \mu\bar{f}(x_2) = x_1. \tag{9.62}$$

Then we have

$$x_1\bar{f}(x_1) = x_2\bar{f}(x_2) = c, \tag{9.63}$$

$$\mu = \frac{x_1x_2}{c}. \tag{9.64}$$

Conversely, if $0 < x_1 < x_2$, with x_1 and x_2 satisfying (9.60), and μ is defined by (9.64), then x_1 and x_2 also satisfy (9.62).

Define κ to be the derivative

$$\kappa = (f^2)'(x_1) = f'(x_1)f'(x_2) = \mu^2\bar{f}'(x_1)\bar{f}'(x_2)$$

occurring in Theorem 7.3. A short calculation gives

$$\kappa = u(x_1)u(x_2), \tag{9.65}$$

where $u(x)$ is the function

$$u(x) = \frac{x\bar{f}'(x)}{\bar{f}(x)}. \tag{9.66}$$

Our basic idea is to use c as a parameter and to express x_1, x_2, μ, and κ as functions of c. To make this rigorous, assume that $\bar{f}:[0,\infty)\to[0,\infty)$ is continuous and C^2 on $(0,\infty)$ and that, if $g(x)=x\bar{f}(x)$, then there exists a number $s_0>0$ such that

$$g'(x)>0 \quad \text{for } 0<x<s_0, \qquad g'(x)<0 \quad \text{for } x>s_0. \tag{9.67}$$

For simplicity in the statement of our theorems, further assume that

$$\lim_{x\to\infty} g(x)=0. \tag{9.68}$$

Define c_* by

$$c_*=g(s_0)=\max_{x>0} g(x), \tag{9.69}$$

and $g_1=g\,|\,[0,c_*]$ and $g_2=g\,|\,[c_*,\infty)$. Then for $0<c\leqq c_*$ (9.63) and the condition $0<x_1<x_2$ determine x_1 and x_2 as functions of c:

$$x_1=x_1(c)=g_1^{-1}(c)\in(0,s_0] \quad \text{and} \quad x_2=x_2(c)=g_2^{-1}(c)\in[s_0,\infty), \tag{9.70}$$

where g_1^{-1} and g_2^{-1} are the inverse functions of g_1 and g_2, respectively. Because $g_1'(x)>0$ for $0<x<s_0$ and g_1 is continuous on $[0,s_0]$, we obtain that x_1 is continuous on $[0,c_*]$ and C^1 on $(0,c_*)$, $x_1(0)=0$, $x_1(c_*)=s_0$ and $x_1'(c)>0$ for $0<c<c_*$. Similarly, we find that x_2 is continuous on $(0,c_*]$ and C^1 on $(0,c_*)$, $x_2(c_*)=s_0$ and $x_2'(c)<0$ for $0<c<c_*$. Note that (9.68) ensures that the domain of x_2 is $(0,c_*]$ and $\lim_{c\to0^+}x_2(c)=\infty$. Having defined $x_1(c)$ and $x_2(c)$, we then have that $\mu=\mu(c)$ and $\kappa=\kappa(c)$, given by (9.64) and (9.65), respectively, are functions of c. To make further progress we must establish some of the properties of $\mu(c)$ and $\kappa(c)$.

LEMMA 9.2. *Assume $\bar{f}:[0,\infty)\to[0,\infty)$ is continuous and C^2 on $(0,\infty)$. If $g(x)=x\bar{f}(x)$ assume there exists $s_0>0$ such that $g'(x)>0$ for $0<x<s_0$ and $g'(x)<0$ for $x>s_0$ and $\lim_{x\to\infty}g(x)=0$. Let $x_1(c)$ and $x_2(c)$ be as defined before for $0<c\leqq c_*=g(s_0)$, and let $\kappa(c)$ and $\mu(c)$ be defined by* (9.64) *and* (9.65). *Then $\mu(c)$ and $\kappa(c)$ have the following properties*:

(i) *$\mu(c)\to s_0^2/c_*>0$ and $\kappa(c)\to1$ as $c\to c_*$.*

(ii) *$\mu(c)\to\infty$ as $c\to0^+$. If $u(x)$ is defined by equation* (9.66), *assume that $\lim_{x\to0+}u(x)=L_1$, where L_1 is finite, and that $\lim_{x\to\infty}u(x)=L_2$, where we allow $L_2=-\infty$. Then we have $\lim_{c\to0^+}\kappa(c)=L_1L_2$.*

(iii) *Suppose that there exists $\theta\geqq0$ such that $\bar{f}'(x)>0$ for $0<x<\theta$ and $\bar{f}'(x)<0$ for $x>\theta$ (so $\theta<s_0$), and that $u'(x)<0$ for $x>\theta$. Define $v(x)=-u(x)$ for $x\geqq\theta$ and let $v^{-1}(\gamma)$ denote the inverse map. Note that $\lim_{x\to\infty}v(x)=-L_2>1$ under our assumptions, define $\delta=\max(-L_2^{-1},-L_1)$, and assume that*

$$g(v^{-1}(\gamma))>g(v^{-1}(1/\gamma)) \quad \text{for } \delta<\gamma<1. \tag{9.71}$$

Then we have $\kappa(c)<1$ for $0<c<c_*$. If we define $\Phi(t)=\log(g(v^{-1}(t)))$ for $\delta<t<1$, inequality (9.71) *will be satisfied if*

$$\Phi'(t)<-\left(\frac{1}{t^2}\right)\Phi'\left(\frac{1}{t}\right) \quad \text{for } \delta<t<1. \tag{9.72}$$

(iv) *If $\kappa(c)<1$ for $0<c<c_*$, then $\mu'(c)<0$ for $0<c<c_*$.*

(v) *If $u'(x)<0$ for all $x>0$ and if $c\in(0,c_*)$ is such that $\kappa(c)\leqq0$, then $\kappa'(c)>0$.*

Proof. (i) By definition of s_0, $g'(s_0)=f(s_0)+s_0f'(s_0)=0$, so we obtain $u(s_0)=-1$. Since we have already noted that $x_1(c)$ and $x_2(c)$ approach s_0 as $c\to c_*$, we conclude that

$$\lim_{c\to c_*}\kappa(c)=\lim_{c\to c_*}u(x_1)u(x_2)=(-1)^2=1.$$

Using the same sort of reasoning, we obtain

$$\lim_{c\to c_*}\mu(c)=\lim_{c\to c_*}\frac{x_1x_2}{c}=\frac{s_0^2}{c_*}.$$

(ii) Because $\bar f$ is continuous at zero, there exists a constant M such that

$$Mx\geqq g(x)\quad\text{for } x \text{ small and positive.}$$

For $c>0$ and small it follows that

$$Mx_1(c)\geqq g(x_1(c))=c\quad\text{or}\quad x_1\geqq c/M.$$

Using this estimate, we see that for $c>0$ small,

$$\mu(c)\geqq(1/M)x_2(c),$$

and because $x_2(c)\to\infty$ as $c\to 0^+$, $\mu(c)\to\infty$ as $c\to 0^+$. Because $\lim_{c\to 0^+}x_1(c)=0$ and $\lim_{c\to 0^+}x_2(c)=\infty$, $\lim_{c\to 0^+}\kappa(c)=\lim_{x\to 0}u(x)\lim_{x\to\infty}u(x)=L_1L_2$.

(iii) Assume that u, f, and g satisfy the given assumptions but that there exists c, $0<c<c_*$, such that $\kappa(c)\geqq 1$. Our assumptions imply $L_1\geqq 0$ and $L_2\leqq 0$, so $\lim_{c\to 0^+}\kappa(c)\leqq 0$. Thus, by choosing a different number c, $0<c<c_*$, we can assume that $\kappa(c)=1$. Because $u(x_2)<\mu(\sigma)=-1$ (since $x_2(c)>\sigma$) and because $u(x)>0$ for $x<\theta$, we must have $x_1(c)\geqq\theta$. If $\theta>0$, so $u(\theta)=0$, then we must in fact have $x_1(c)>0$ and of course $x_1(c)>\theta$ if $\theta=0$. If we write $\gamma=v(x_1(c))$, then we must have $0=v(\theta)<\gamma<1=v(s_0)$ if $\theta>0$ and $-L_1<\gamma<1$ if $\theta=0$. Furthermore, we must have $\gamma^{-1}=v(x_2(c))$, so $\gamma^{-1}\in(1,-L_2)$, and we obtain the estimates $\delta<\gamma<1$. Now (9.63) gives

$$g(v^{-1}(\gamma))=g(v^{-1}(1/\gamma)),$$

which contradicts (9.71).

Note that (9.71) is equivalent to

$$\Phi(1)-\Phi(\gamma)<\Phi(1)-\Phi(\gamma^{-1})\quad\text{for } \delta<\gamma<1. \tag{9.73}$$

Using the fundamental theorem of calculus and changing variables in the integral for the right-hand side, we see that (9.73) is equivalent to

$$\int_\gamma^1\Phi'(t)\,dt<\int_\gamma^1-\left(\frac{1}{t^2}\right)\Phi'\left(\frac{1}{t}\right)dt,$$

so (9.72) implies (9.73) and (9.71).

(iv) A calculation shows

$$(\mu(c))^2=\frac{x_1x_2}{\bar f(x_1)\bar f(x_2)}\quad\text{where } x_j=x_j(c), \tag{9.74}$$

so it suffices to show that if $\kappa(c)<1$, then

$$\frac{d}{dc}\log\left(\frac{x_1x_2}{\bar f(x_1)\bar f(x_2)}\right)<0. \tag{9.75}$$

Using the formula $x_j'(c)=(\bar f(x_j)^{-1}(1+u(x_j)))^{-1}$ (which we obtain by differentiating $x_j\bar f(x_j)=c$), we find that (9.75) is equivalent to

$$\sum_{j=1}^2(x_j\bar f(x_j)(1+u(x_j)))^{-1}-\sum_{j=1}^2\bar f'(x_j)(\bar f(x_j)^2(1+u(x_j)))^{-1}<0.$$

Multiplying the above inequality by $x_1x_2\bar{f}(x_1)\bar{f}(x_2)=c^2$ and simplifying, we find that the above inequality is equivalent to

$$2c(1+u(x_1))^{-1}(1+u(x_2))^{-1}(1-\kappa(c))<0. \tag{9.76}$$

Recall that $g'(x)>0$ for $0<x<s_0$ and $g'(x)<0$ for $x>s_0$, which implies $u(x)>-1$ for $0<x<s_0$ and $u(x)<-1$ for $x>s_0$. From this we conclude that $(1+u(x_1))(1+u(x_2))$ is negative and that $\mu'(c)<0$ if and only if $\kappa(c)<1$.

(v) If $\kappa(c)\leqq 0$, we must have $u(x_1(c))\geqq 0$, since $u(x_2(c))<-1$ for $0<c<c_*$. It follows that

$$\kappa'(c)=u'(x_1)x_1'u(x_2)+u(x_1)u'(x_2)x_2'>0,$$

because we are assuming that $u'(x)<0$ for all $x>0$, and that $x_1'(c)>0$ and $x_2'(c)<0$ for $0<c<c_*$. □

With $j=4$ or 5 and parameters $\nu>0$ and $\lambda>\nu+1$, let $\bar{f}_j$ be defined as in (9.24) and let functions $\mu_j(c)$ and $\kappa_j(c)$ be defined by substituting $\bar{f}_j$ for $\bar{f}$ in (9.63)-(9.65). Define $g_j(x)=x\bar{f}_j(x)$, so that $g_j'(x)>0$ for $0<x<\sigma_j$ (where σ_j is given in Theorem 9.3), and $g_j'(x)<0$ for $x>\sigma_j$. Also recall that $\bar{f}_j'(x)>0$ for $0<x<\theta_j$ and $\bar{f}_j'(x)<0$ for $x>\theta_j$, where θ_j is an in Theorem 9.3. From Lemma 9.2 we obtain Lemma 9.3.

LEMMA 9.3. *With $j=4$ or 5 and parameters $\nu>0$ and $\lambda>\nu+1$, the functions $\mu_j(c)$ and $\kappa_j(c)$ (defined for $0<c\leqq c_{*j}=g_j(\sigma_j)$) have the following properties*:

(i) $u_j(c)\to\sigma_j^2/c_{*j}>0$ *and* $\kappa_j(c)\to 1$ *as* $c\to c_{*j}$.
(ii) $\mu_j(c)\to\infty$, $\kappa_4(c)\to-\infty$ *and* $\kappa_5(c)\to-\nu(\lambda-\nu)<0$ *as* $c\to 0^+$.
(iii) $\kappa_j(c)<1$, *if* $0<c<c_{*j}$.
(iv) $\mu_j'(c)<0$, *if* $0<c<c_{*j}$.
(v) $\kappa_j'(c)>0$, *if* $0<c<c_{*j}$.

Proof. Define $u_j(x)=x\bar{f}_j'(x)(\bar{f}_j(x))^{-1}$ and $v_j(x)=-u_j(x)$. A calculation gives

$$u_4(x)=\nu-x, \qquad u_5(x)=[\nu-(\lambda-\nu)x^\lambda](1+x^\lambda)^{-1},$$

so $u_j'(x)<0$ for all $x>0$. A further calculation gives

$$v_4^{-1}(t)=\nu+t, \qquad (v_5^{-1}(t))^\lambda=(\nu+t)(\lambda-\nu-t)^{-1}.$$

Lemma 9.2 will imply Lemma 9.3 if we can prove that (9.71) is satisfied with g_j and v_j replacing g and v, and Lemma 9.2 implies that this will be the case if

$$\Phi_j'(t)<-\left(\frac{1}{t^2}\right)\Phi_j'\left(\frac{1}{t}\right) \quad \text{for } \delta_j<t<1, \tag{9.77}$$

where $\Phi_j(t)=\log g_j(v_j^{-1}(t))$ and $\delta_4=0$ and $\delta_5=(\lambda-\nu)^{-1}$. We can easily check that

$$\Phi_4'(t)=(1-t)(\nu+t)^{-1}, \qquad \Phi_5'(t)=(1-t)(\nu+t)^{-1}(\lambda-\nu-t)^{-1},$$

so for $j=4$ inequality (9.77) is equivalent to

$$(1-t)(\nu+t)^{-1}<(1-t)(\nu t^3+t^2)^{-1} \quad \text{for } 0<t<1 \tag{9.78}$$

and for $j=5$ inequality (9.77) is equivalent to

$$\begin{aligned}&(1-t)(\nu+t)^{-1}(\lambda-\nu-t)^{-1}<(1-t)t^{-1}(\nu t+1)^{-1}(\lambda t-\nu t-1)^{-1}\\ &\text{for } (\lambda-\nu)^{-1}<t<1.\end{aligned} \tag{9.79}$$

Since $0<t<1$, (9.78) is obviously true. Because $0<t<1$ we have $\nu+t>\nu t^2+t>0$;

and by using the fact that $\lambda > \nu$ we can see

$$\lambda - \nu - t > \lambda t - \nu t - 1 > 0 \quad \text{for } (\lambda - \nu)^{-1} < t < 1,$$

so inequality (9.79) is valid. □

With the aid of Lemma 9.2 we can give conditions under which a function $\mu\bar{f}(x)$ satisfies (II) precisely for $\mu \in (\mu_0, \sigma]$, where $\sigma > \mu_0$.

THEOREM 9.6. *Assume $\bar{f}:[0,\infty)\to[0,\infty)$ is a continuous map that is C^3 on $(0,\infty)$. Assume there exists $\theta \geqq 0$ such that $\bar{f}'(x) > 0$ for $0 < x < \theta$ and $\bar{f}'(x) < 0$ for $x > \theta$. If $g(x) = x\bar{f}(x)$, assume there exists $s_0 > 0$ such that $g'(x) > 0$ for $0 < x < s_0$ and $g'(x) < 0$ for $x > s_0$. Define $u(x) = x\bar{f}'(x)(\bar{f}(x))^{-1}$ and $v(x) = -u(x)$ and assume $u'(x) < 0$ for all $x > 0$ and*

$$g(v^{-1}(\gamma)) > g(v^{-1}(1/\gamma)) \quad \text{for } \delta < \gamma < 1, \tag{9.80}$$

where δ is defined as in Lemma 9.2 and v^{-1} is the inverse map of v. (Recall that inequality (9.80) is satisfied if inequality (9.72) holds.) Finally, suppose there exists a C^3 map ψ of an interval (a, b) onto $(0,\infty)$ such that $\psi'(x) > 0$ for $x \in (a, b)$ and $\psi^{-1}(\mu\bar{f})\psi$ has negative Schwarzian derivative for all $x \in (a, b)$. Then there exists $\sigma > \mu_0 = \bar{f}(s_0)s_0^{-1}$ such that $\mu\bar{f}(x)$ satisfies (II) at $x_0(\mu) = x_0$, the unique fixed point of $\mu f(x)$ in the interval (θ,∞), if and only if $\mu_0 < \mu \leqq \sigma$. If $\mu(c)$ and $\kappa(c)$ are defined as in Lemma 9.2 and $\kappa(c_2) = -1$, then $\mu(c_2) = \sigma$.

Proof. If $f = \mu\bar{f}$ is as in Theorem 7.3 or 7.4, but we assume that $\psi^{-1}f\psi$ (ψ as in Theorem 9.6) has negative Schwarzian derivative everywhere instead of supposing that f has, we can still easily see (using the remarks at the beginning of this section) that the conclusions of Theorem 7.3 and 7.4 are satisfied.

Now let $\mu(c)$ and $\kappa(c)$ be as in Lemma 9.2. Theorem 9.3 implies that $\mu\bar{f}(x)$ satisfies condition (0) at $x_0(\mu)$ if and only if $\mu > \mu_0 = \bar{f}(s_0)s_0^{-1}$, and Lemma 9.2 implies that $\mu_0 = \mu(c_*)$ and that $\mu(c) > \mu_0$ for $0 < c < c_*$. Since (Lemma 9.2) $\mu'(c) < 0$ for $0 < c < c_*$ and $\lim_{c\to 0^+}\mu(c) = \infty$, we will work with the parameter c instead of $\mu > \mu_0$. Define a number c_1 by

$$c_1 = \inf\{c > 0: \mu(\gamma)\bar{f}(x) \text{ satisfies (I) at } x_0 \text{ for } c \leqq \gamma < c_*\}.$$

Corollary 7.1 implies that $c_1 < c_*$. Define c_2 as in the statement of the theorem if $\kappa(c) = -1$ has a solution $c > 0$; otherwise define $c_2 = 0$. Lemma 9.2 implies that $\kappa(c) < 1$ for $0 < c < c_*$ and $\kappa(c) \geqq -1$ if and only if $c \geqq c_2$. Thus Theorem 7.3 will imply that $\mu(c)\bar{f}(x)$ satisfies (II) if and only if $c_2 \leqq c < c_*$ if we can prove that $c_2 > c_1$ when $c_1 > 0$. However, if $c_1 > 0$, Theorem 7.4 and Corollary 7.2 imply that there exists $\delta < 0$ such that $\mu(c_1) \leqq \mu < \mu(c_1) + \delta$, $(\mu\bar{f})^2$ has a fixed point x such that $(d/dx)(\mu\bar{f})^2(x) < -1$. If $c_2 \leqq c_1$, this contradicts Lemma 9.2, so we must have $c_2 > c_1$. □

As an immediate consequence of Theorem 9.6 and Lemma 9.3 we obtain Corollary 9.1.

COROLLARY 9.1. *For parameters $\nu \geqq 1$ and $\lambda > \nu + 1$ let $\bar{f}_4(x)$ and $\bar{f}_5(x)$ be as defined before and let $\mu_0(\nu)$ and $\mu_0(\nu, \lambda)$ be as defined in Table 1 for the functions $\bar{f}_4$ and $\bar{f}_5$, respectively. If $\bar{f}_j$ has its maximum on $(0,\infty)$ at θ_j, there exist continuous functions $\sigma_4(\nu)$ and $\sigma_5(\nu, \lambda)$ such that μf_j satisfies condition (II) at the unique fixed point of μf_j in (θ_j,∞) if and only if $\mu_0(\nu) < \mu \leqq \sigma_4(\nu)$ for $j = 4$ or $\mu_0(\nu, \lambda) < \mu \leqq \sigma_5(\nu, \lambda)$ for $j = 5$.*

Proof. We need prove the continuity and finiteness of σ_j, and this follows easily from the results of Lemma 9.3. □

If $\psi:(a, b)\to(0,\infty)$ is as in Theorem 9.6, the results of this section also apply to the functions $\psi^{-1}(\mu\bar{f}_j)\psi$ for $\nu \geqq 1$ and $\lambda > \nu + 1$. Taking $\psi(x) = ax$, $a > 0$, we obtain, for example, the conclusion of Corollary 9.1 for $\mu_1 x^\nu e^{-ax}$ and $\mu_1 x^\nu(1 + bx^\lambda)^{-1}$, where

$\mu_1 > 0$ and $b = a^\lambda$ is an arbitrary positive number. Taking $\psi(x) = x^p$ for $p > 0$, we obtain the same results for $\mu_2 x^\nu e^{-bx^p}$ and $\mu_2 x^\nu (1+bx^{\lambda p})^{-1/p}$, where $\mu_2 > 0$, $b > 0$, and $p > 0$.

Although we will not pursue this here, we can establish the conclusions of Corollary 9.1 for other classes of functions, e.g., $\mu \bar{f}_6(x)$, where

$$\bar{f}_6(x) = x^\nu \exp(-x(1+ax)),$$

where $\nu \geqq 1$ and $a > 0$. The major problem is verifying (9.71) or (9.72).

REFERENCES

[1] D. J. ALLWRIGHT, *Hypergraphic functions and bifurcation in recurrence relations*, SIAM J. Appl. Math., 34 (1978), pp. 687-691.

[2] S. P. BLYTHE, R. M. NISBET, AND W. S. C. GURNEY, *Instability and complex dynamic behaviour in population models with long time delays*, Theoret. Population Biol., 2 (1982), pp. 147-176.

[3] S. CHAPIN AND R. D. NUSSBAUM, *Asymptotic estimates for the periods of periodic solutions of a differential-delay equation*, Michigan Math. J., 31 (1984), pp. 215-229.

[4] S.-N. CHOW AND D. GREEN JR., *Some results on singular delay-differential equations*, in Chaos. Fractals and Dynamics, Vol. 1, W. R. Smith and P. Fisher, eds., Marcel Dekker, New York, 1985, pp. 161-182.

[5] S.-N. CHOW AND J. MALLET-PARET, *Singularly perturbed delay-differential equations*, in Coupled Nonlinear Oscillators, J. Chandra and A. C. Scott, eds., North-Holland Math. Studies, Vol. 80, 1983, pp. 7-12.

[6] P. COLLET AND J.-P. ECKMANN, *Iterated Maps on the Intervals as Dynamical Systems*, Birkhäuser, Boston, 1980.

[7] M. W. DERSTINE, H. M. GIBBS, F. A. HOPF, AND D. L. KAPLAN, *Alternate paths to chaos in optical bistability*, Phys. Rev. A(3), 27 (1983), pp. 3200-3208.

[8] H. M. GIBBS, F. A. HOPF, D. L. KAPLAN, AND R. L. SHOEMAKER, *Observation of chaos in optical bistability*, Phys. Rev. Lett., 46 (1981), pp. 474-477.

[9] W. S. C. GURNEY, S. P. BLYTHE, AND R. M. NISBET, *Nicholson's blowflies revisited*, Nature, 287 (1980), pp. 17-21.

[10] K. P. HADELER AND J. TOMIUK, *Periodic solutions of difference-differential equations*, Arch. Rational Mech. Anal., 65 (1977), pp. 87-95.

[11] U. AN DER HEIDEN AND M. C. MACKEY, *The dynamics of production and destruction: Analytic insight into complex behaviour*, J. Math. Biol., 16 (1982), pp. 75-101.

[12] U. AN DER HEIDEN, M. C. MACKEY, AND H. O. WALTHER, *Complex oscillations in a simple deterministic neuronal network*, in Mathematical Aspects of Physiology, F. C. Hoppensteadt, ed., Lectures in Applied Mathematics, Vol. 19, American Mathematical Society, Providence, RI, 1981, pp. 355-360.

[13] U. AN DER HEIDEN AND H.-O. WALTHER, *Existence of chaos in control systems with delayed feedback*, J. Differential Equations, 47 (1983), pp. 273-295.

[14] K. IKEDA, *Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system*, Opt. Comm., 30 (1979), pp. 257-261.

[15] G. JULIA, *Memoire sur l'iteration des fonctions rationelles*, J. Math. Pures Appl. (9), 4 (1918), pp. 47-245.

[16] M. C. MACKEY AND L. GLASS, *Oscillation and chaos in physiological control systems*, Science, 197 (1977), pp. 287-289.

[17] M. C. MACKEY AND U. AN DER HEIDEN, *Dynamical diseases and bifurcations: Understanding functional disorders in physiological systems*, Funkt. Biol. Med., 156 (1982), pp. 156-164.

[18] J. MALLET-PARET, *Morse decompositions and global continuation of periodic solutions for singularly perturbed delay equations*, in Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., Reidel, Dordrecht, the Netherlands, 1983, pp. 351-365.

[19] ———, *Morse decompositions for delay-differential equations*, J. Differential Equations, 72 (1988), pp. 270-315.

[20] J. MALLET-PARET AND R. D. NUSSBAUM, *Global continuation and complicated trajectories for periodic solutions of a differential-delay equation*, Proc. Sympos. Pure Math., 45 (1986), pp. 155-167.

[21] ———, *Global continuation and asymptotic behaviour for periodic solutions of a differential-delay equation*, Annali di Matematica Pura ed Appl., 145 (1986), pp. 33-182.

[22] ———, *A bifurcation gap for a singularly perturbed delay equation*, in Chaotic Dynamics and Fractals, Vol. 2, M. Barnsley and S. Demko, ed., Academic Press, New York, 1986, pp. 263-287.

[23] R. D. NUSSBAUM, *Periodic solutions of nonlinear autonomous functional differential equations*, Lecture Notes in Mathematics 730, Springer-Verlag, Berlin, New York, 1979, pp. 283-325.

[24] R. D. NUSSBAUM, *Boundary layer phenomena for a differential-delay equation*, submitted for publication.
[25] ———, *Circulant matrices and differential-delay equations*, J. Differential Equations, 60 (1985), pp. 201–217.
[26] D. SAUPE, *Global bifurcation of periodic solutions to some autonomous differential delay equations*, Appl. Math. Comput., 13 (1983), pp. 185–211.
[27] D. SINGER, *Stable orbits and bifurcation of maps of the interval*, SIAM J. Appl. Math., 35 (1978), pp. 260–267.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 293-307, March 1989

002

SINGULAR SOLUTIONS AND ILL-POSEDNESS FOR THE EVOLUTION OF VORTEX SHEETS*

RUSSEL E. CAFLISCH† AND OSCAR F. ORELLANA‡

Abstract. The evolution of a planar vortex sheet is described by the Birkhoff-Rott equation. Duchon and Robert [*C.R. Acad. Sci. Paris*, 302 (1986), pp. 183-186], [Comm. Partial Differential Equations, 13 (1988), pp. 1265-1295] have constructed exact solutions of this equation that are analytic for all $t<0$ but have a possible singularity in the curvature of the sheet at $t=0$. This shows that smooth initial data for a vortex sheet can lead to singularity formation at a finite time, in agreement with the results of numerical computation [*J. Fluid Mech.*, 167 (1986), pp. 65-93], [*J. Fluid Mech.*, 114 (1982), pp. 283-298] and of asymptotic expansion [*Proc. Roy. Soc. London Ser.* A, 365 (1979), pp. 105-119], [*Theoretical and Applied Mechanics*, in Proc. XVI Internat. Congr. Theoret. Appl. Mech., F. I. Niordson and N. Olhoff, eds., North-Holland, Amsterdam, 1984, pp. 629-633]. We present an independent construction of these solutions and use these results to infer that the vortex sheet problem is ill-posed in Sobolev class H_n with $n>3/2$. Earlier results show well-posedness in an analytic function class [Comm. Pure Appl. Math., 39 (1986), pp. 807-838], [*Comm. Math. Phys.*, 80 (1981), pp. 485-516]. Our method is to construct an explicit singular function that is a solution of the linearized equation, with a correction term added on to make the sum an exact solution of the nonlinear equation. The correction term is analyzed using the Cauchy-Kowalewski method.

Key words. vortex sheets, vorticity, Kelvin-Helmholtz instability, fluid dynamics, Birkhoff-Rott equation, singularities, ill-posedness, instability, Euler equations

AMS(MOS) classifications. 76C05, 96E30, 35L67

1. Introduction.

A planar vortex sheet is a curve in a two-dimensional, inviscid, incompressible flow along which the fluid velocity is discontinuous. Vortex sheets, and the Kelvin-Helmholtz instability that they undergo, play an important role in many flows, such as mixing layers, two-fluid interfaces and flow past airfoils. Asymptotic analysis by Moore [11], [12] and numerical computation by Krasny [8] and Meiron, Baker, and Orszag [10] have shown that a vortex sheet may develop a singularity, i.e., infinite curvature at a point, in a finite time. The appearance of this singularity is important because it is immediately followed by rollup of the sheet [9]. A more mathematical reason for interest in such singularity formation is that it may serve as a simple analogue of singularity formation for the three-dimensional Euler equations.

Duchon and Robert [6, 16] perform a general construction of vortex sheet solutions that are analytic for all $t>0$, by choosing initial data to lie on the stable manifold for the Birkhoff-Rott equation ((1.1) below). The initial data can have singularities in the $(1+\nu)$th derivative for any $\nu>0$ (for a precise statement see [6], [16]; fractional derivatives can be understood in the Hölder sense as in (1.6)). Our aim is to present an independent construction of these singular solutions and to discuss their significance. We also obtain slightly more precise pointwise information on the singularity, by using a pointwise norm rather than the Fourier norm in [6], [16].

The singular solutions found here or in [6], [16] can be used to construct exact solutions for vortex sheets that develop singularities at finite time starting from smooth

* Received by the editors August 17, 1987; accepted for publication (in revised form) June 3, 1988.

† Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR 85-0017 and University Research Initiative grant AFOSR 86-0352 and by the Alfred P. Sloan Foundation.

‡ Universidad Técnica Federico Santa María, Valparaiso, Chile. The research of this author was supported in part by Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) under grant 235 and Universidad Técnica Federico Santa María, Valparaiso, Chile.

initial data, as described in the concluding § 7. This shows that for the two-dimensional Euler equations, singular initial data (a vortex sheet) can become more singular in finite time. Furthermore, the existence of such singular solutions shows the vortex sheet problem to be ill posed in Sobolev space H^n for $n > 3/2$. A somewhat different proof of ill-posedness for vortex sheets was given by Ebin [15].

The vortex sheet is parametrized by a real variable γ defined so that it is Lagrangian (the value of γ is constant on a given fluid particle) and so that the density of circulation with respect to the γ variable on the sheet is 1. The position of the vortex sheet is defined by a complex variable $z(\gamma, t)$ that satisfies the Birkhoff-Rott equation [1]:

$$\frac{\partial}{\partial t}\bar{z}(\gamma, t) = \frac{1}{2\pi i} PV \int_{-\infty}^{\infty} \frac{d\gamma'}{z(\gamma, t) - z(\gamma', t)}, \tag{1.1}$$

in which the integral is a Cauchy principal value. An arbitrary constant in the definition of the Cauchy integral is irrelevant since the right-hand side involves a difference. Moreover, we will assume for simplicity that z is odd in γ, i.e.,

$$z(-\gamma, t) = -z(\gamma, t) \tag{1.2}$$

so that $z(0, t) = 0$. The function $z = \gamma$ is a steady solution of (1.1) and corresponds to a flat vortex sheet with a uniform density of circulation.

Linearization of (1.1) about the steady solution $z = \gamma$ yields the following equation for $z = \gamma + s$:

$$\partial_t \bar{s}(\gamma, t) = \tfrac{1}{2} H[s_\gamma] = \tfrac{1}{2}(s_{+\gamma} - s_{-\gamma}) \tag{1.3}$$

in which H is the Hilbert transform defined by $H[s] = s_+ - s_-$, s_+ is the upper analytic part of s, i.e., the part with positive Fourier wavenumbers, and s_- is the lower analytic part. The linearized equation (1.3) is unstable with modes $e^{ik\gamma \pm kt/2}$. In this paper we are investigating the nonlinear behavior of this Kelvin-Helmholtz instability. Since the linearized modes have arbitrarily large temporal growth rates, we find that singularities may develop in finite time for the solutions of the nonlinear equation.

An explicit example of our singular solutions is

$$z(\gamma, t) = \gamma + s_0 + r, \tag{1.4}$$

$$s_0(\gamma, t) = \varepsilon(1-i)\{(1 - e^{-t/2 - i\gamma})^{1+\nu} - (1 - e^{-t/2 + i\gamma})^{1+\nu}\} \tag{1.5}$$

in which ε is small. The dominant term s_0 is an exact solution of the linearized equation (1.3) and the correction term r is negligible relative to s_0 as explained in Theorem 1. Since $s_0 \simeq c\gamma^{1+\nu}$ for $t = 0$ and $\gamma \simeq 0$, then $z_{\gamma\gamma} \simeq s_{0\gamma\gamma} \simeq c\gamma^{\nu-1}$ (with a different constant c). Therefore the vortex sheet has an infinite curvature at $\gamma = 0$, $t = 0$ for $0 < \nu < 1$. For the approximate singular solution $z_0 = \gamma + s_0$, with s_0 given by (1.5), the vortex strength $|z_{0\gamma}|^{-1}$ $(t = 0)$ is plotted in Fig. 1. The cusp at $\gamma = 0$ is due to the singularity. An explanation of the terms in (1.5) is as follows. The $e^{\pm i\gamma}$ makes s_0 periodic and odd. The exponent $-t$ gives a solution that decays in time. Such a decaying solution requires the factors $(1-i)$ and $\frac{1}{2}$.

To define a more general class of functions s_0, let ε be a small real number and let $1 > \nu > \alpha > 0$. For s analytic in $|\mathrm{Im}\ \gamma| < \rho$ define the Hölder norm

$$|s|_\rho = \sup_{|\mathrm{Im}\ \gamma| < \rho} |s(\gamma)| + \sup_{\substack{|\mathrm{Im}\ \gamma|, |\mathrm{Im}\ \gamma'| < \rho \\ \gamma \neq \gamma'}} \frac{|s(\gamma) - s(\gamma')|}{|\gamma - \gamma'|^\alpha}. \tag{1.6}$$

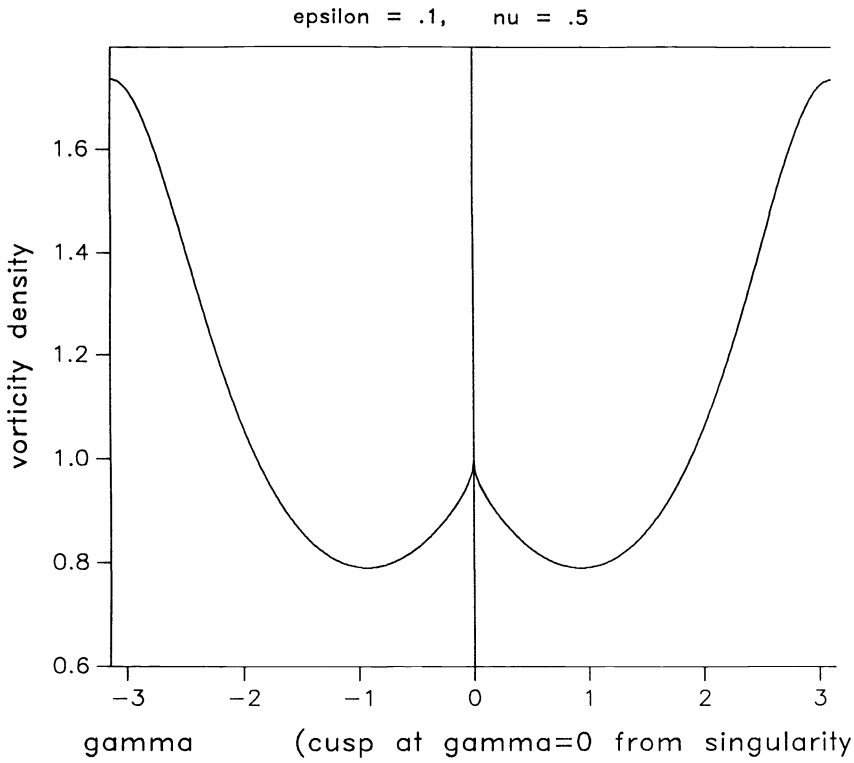


FIG. 1. *Approximate vorticity density* $\sigma = |z_{0\gamma}|^{-1} = |1 + s_{0\gamma}|^{-1}$ *with* s_0 *defined by* (1.5), $\varepsilon = 0.1$, $\nu = 0.5$. *The cusp at* $\gamma = 0$ *corresponds to the singularity.*

We require s_0 to satisfy the following:

(i) s_0 solves the linearized equation (1.3);

(ii) s_0 is analytic in the time-dependent strip $|\text{Im } \gamma| < |t|/2$ for $t > 0$;

(iii) s_0 is small and decays to zero as $t \to \infty$. At $t = 0$ s_0 has (at most) a singularity in its $(1+\nu)$th derivative, i.e.,

$$|s_0|_\rho + |s_{0\gamma}|_\rho < c\varepsilon\, e^{-(|t|/2-\rho)}, \tag{1.7}$$

$$|s_{0\gamma\gamma}|_\rho < c\varepsilon(1+(|t|-2\rho)^{\nu-\alpha-1})\, e^{-(|t|/2-\rho)} \tag{1.8}$$

for $t > 0$.

Note that the function s_0 in (1.5) satisfies (i)-(iii). Our main result is the following existence theorem.

THEOREM 1. *Let* ε *be a sufficiently small real number and let* $1 > \nu > \alpha > 0$. *Let* s_0 *satisfy* (i)-(iii) *above*; *i.e.*, s_0 *is an analytic solution of the linearized equation* (1.3) *that decays to zero at* $t = \infty$ *and has a mild singularity at* $t = 0$. *Then there is a function* $r(\gamma, t)$ *such that*

$$z(\gamma, t) = \gamma + s_0 + r \tag{1.9}$$

is an analytic solution of the Birkhoff-Rott equation (1.1) *for* $t > 0$ *and* $\kappa|\text{Im } \gamma| < t$ *in which* $\kappa > 2$ *and* $\kappa \to 2$ *as* $\varepsilon \to 0$. *Moreover, r can be chosen so that the decaying mode* $r_+ + i\bar{r}_- = 0$ *at* $t = 0$ *and that*

$$|r|_0 + |r_\gamma|_0 < c\varepsilon^2 \exp(-|t|/2), \tag{1.10}$$

$$|r_{\gamma\gamma}|_0 < c\varepsilon^2(1+|t|^{\nu-\alpha-1}) \exp(-|t|/2) \tag{1.11}$$

in which c is some constant that is independent of ε *and depends smoothly on* α^{-1} *and* $(\nu-\alpha)^{-1}$ (*i.e., c may be infinite at* $\alpha = 0$ *or* $\alpha = \nu$).

Since $r_+ + i\bar{r}_- = 0$ at $t = 0$, in some sense half of the initial data of z is given by $\gamma + s_0$. The norm in (1.10), (1.11) is that of (1.6) with $\rho = 0$. These estimates show that r is as smooth as s_0, but much smaller. Stronger estimates for r on the strip $\kappa|\text{Im } \gamma| < t$ are given at the end of § 6.

There are three main ideas in this construction: The first is to extend the Birkhoff-Rott equation to complex γ. Linearization (or equivalently the form of the Kelvin-Helmholtz instability) shows the Birkhoff-Rott equation to be approximately hyperbolic in the imaginary γ direction, so that singularities will move in that direction. This was first pointed out by Moore [11]. Of course, only real values of γ are physically meaningful; i.e., singularities are physically observable only when they lie on the real γ line.

The second idea is to put the singularity in the initial data. By proper choice of initial data, the singularities can be expected to travel approximately on the lines Im $\gamma = \pm t/2$. Although our method cannot track these singularities exactly, we are able to show that the resulting solution is analytic in the wedge $|\text{Im } \gamma| < \kappa t/2$, $t > 0$, as shown in Fig. 2. Note that for all $t > 0$ this wedge contains the line Im $\gamma = 0$. Therefore the vortex sheet is analytic for all $t > 0$.

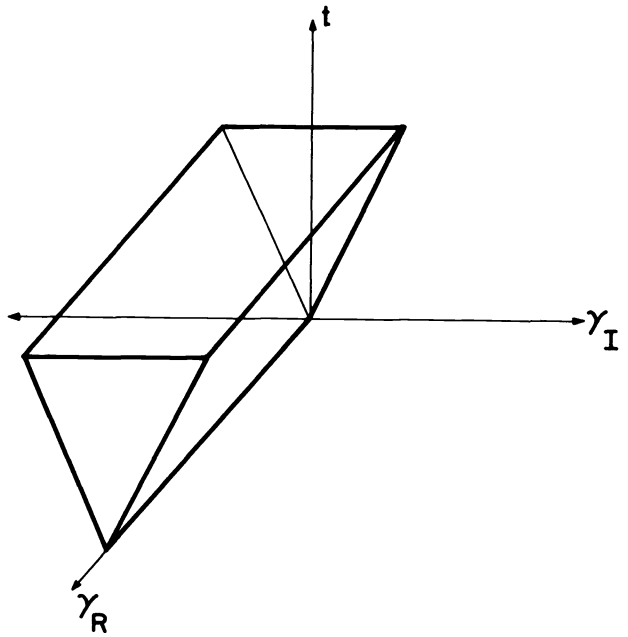


FIG. 2. *Domain of analyticity (in γ) for $z(\gamma, t)$ (equivalently for $r(\gamma, t)$)*: $\{|\text{Im } \gamma| < ct, t > 0\}$.

The third idea is to construct the solution within the class of analytic functions. In this function class the Birkhoff-Rott equation has been shown to be well posed; while in almost any larger class it is expected to be ill posed. For the Sobolev spaces, ill-posedness is shown in § 7. Use of analytic functions provides the stabilization necessary to construct exact solutions in the presence of physical instability. Aside from this practical justification, our belief that the imposition of analyticity is consistent with the zero viscosity limit justifies the use of analytic functions, at least for some flow regimes.

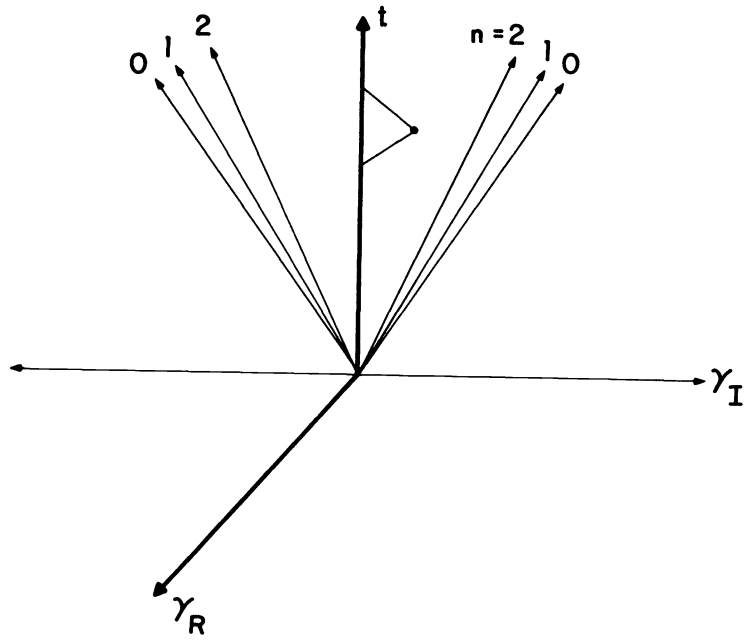


FIG. 3. *Domain of analyticity for the iterates* r_n *(or equivalently* p_n, q_n*). The width of the domain decreases as n increases. The line segments emanating from the plane* $\{\text{Im } \gamma = 0\}$ *are the characteristics used to extend the iterates to complex* γ*, as in* (2.12), (2.13), *or* (3.3).

The correction term r of Theorem 1 is constructed using the Cauchy-Kowalewski method, somewhat modified from that in [13] involving an iteration. At each iteration in our construction, the approximate solution is analytic in a wedge $|\text{Im } \gamma| < ct$, with an angle c that is slightly reduced at each step, as indicated in Fig. 3. The final solution is valid in a limiting wedge, which is shown to be nonempty.

Other analytic results for vortex sheets include a proof of short-time existence by Sulem, Sulem, Bardos, and Frisch and a proof of existence almost up to the expected time of singularity formation by Caflisch and Orellana [2]. The singularities described in the present paper are not accompanied by concentrations of energy. Thus they are much weaker than those discussed by DiPerna and Majda [3]-[5], which possibly appear on a vortex sheet at a later time. On the other hand, there are no known examples for which such energy concentrations develop from less singular initial data.

The outline of this paper follows. Section 2 contains a reformulation of (1.1) and shows the sense in which the problem is hyperbolic. In particular, an equation for the error term r is derived. In § 3 we describe an iteration method for solving the equation for r. Each iteration involves solving an elliptic problem in t and Re γ, then extending the solution to complex γ by solving a hyperbolic problem. The elliptic problem is solved using a Green's function; we solve the hyperbolic problem by integrating along characteristics. Estimates on the first iterate are obtained in § 4, the induction method is described in § 5 and estimates for the successive iterates are derived using a Cauchy-Kowalewski method in § 6. At the end of that section, we summarize the proof of Theorem 1, and we derive two consequences of Theorem 1 in the concluding section.

2. Formulation. We write z as a perturbation of the flat sheet by $z = \gamma + s$ and assume that s is 2π-periodic, i.e.,

$$s(\gamma + 2\pi, t) = s(\gamma, t). \tag{2.1}$$

The Birkhoff-Rott equation (1.1) is equivalent to the following equation for s, which is now written in a way that is analytic for complex γ,

$$\frac{\partial}{\partial t} s^*(\gamma, t) = B[s] \equiv \frac{1}{2\pi i} PV \int_{-\infty}^{\infty} \frac{d\zeta}{-\zeta + s(\gamma) - s(\gamma+\zeta)}, \tag{2.2}$$

in which s^* is the analytic extension of $\bar{s}$, i.e., $s^*(\gamma) = \overline{s(\bar{\gamma})}$. The operator B is expanded as $B[s] = B_1[s] + B_r[s]$ in which

$$B_1[s](\gamma) = -\frac{1}{2\pi i} PV \int_{-\infty}^{\infty} \frac{s(\gamma) - s(\gamma+\zeta)}{\zeta^2} d\zeta = \frac{1}{2} H[s_\gamma] \tag{2.3}$$

is the linear part of B and B_r is the nonlinear remainder. The linearized equation (1.3) is analytically extended as $s_t^* = B_1[s]$.

Since s_0 is an exact solution of this linearized equation, the Birkhoff-Rott equation (1.1) for $z = \gamma + s_0 + r$ can be rewritten as follows:

$$r_t^* = B_1[r] + B_r[s_0 + r]. \tag{2.4}$$

Define the decaying component p and growing component q for r as follows:

$$p = r_+ + i(r_-)^*, \qquad q = r_+ - i(r_-)^*. \tag{2.5}$$

Then (2.4) can be rewritten as

$$p_t = \frac{i}{2} p_\gamma + a, \tag{2.6}$$

$$q_t = \frac{-i}{2} q_\gamma + b, \tag{2.7}$$

in which

$$a = (B_-)^* + iB_+, \qquad b = (B_-)^* - iB_+ \tag{2.8}$$

and B_+ and B_- are the upper and lower analytic components of $B_r[s_0 + r]$. Note that a and b, and thus also p and q, have only components with positive wave numbers.

The system (2.6), (2.7) is elliptic in γ, t for real γ, but it is hyperbolic in γ, t for imaginary γ (or more precisely, in the imaginary γ direction).

We solve (2.6), (2.7) in two parts. For γ real, $t \geqq 0$, we solve the elliptic equation using a Green's function. For $\operatorname{Im} \gamma \neq 0$, $t \geqq 0$, we solve the hyperbolic equation (2.4) by integration along characteristics using the values of r on $\operatorname{Im} \gamma = 0$ as "initial values." These characteristics are indicated in Fig. 3.

We require p and q to be 2π-periodic, to vanish at $t = \infty$, and to have only components with positive wave numbers and $p = 0$ at $t = 0$; i.e.,

$$\begin{gathered} (p, q)(\gamma, t) = (p, q)(\gamma + 2\pi, t), \\ (p, q) \to 0 \quad \text{as } t \to \infty, \qquad p(t=0) = 0, \\ (\hat{p}, \hat{q})(k, t) = \int_0^{2\pi} e^{-ik\gamma} (p, q)(\gamma, t)\, d\gamma = 0 \quad \text{for } k \leqq 0. \end{gathered} \tag{2.9}$$

Under these conditions and for a and b having only components with positive wave numbers, the solution of (2.6), (2.7) is uniquely given by

$$p(\gamma, t) = \int_0^t \int_0^{2\pi} g(\gamma', t - t') a(\gamma - \gamma', t')\, d\gamma'\, dt', \tag{2.10}$$

$$q(\gamma, t) = \int_t^{\infty} \int_0^{2\pi} g(\gamma', t - t') b(\gamma - \gamma', t')\, d\gamma'\, dt' \tag{2.11}$$

for γ real and by

$$p(\gamma+i\mu, t)=p(\gamma, t+2\mu)-2\int_0^\mu a(\gamma+i\mu', t+2(\mu-\mu'))\,d\mu', \tag{2.12}$$

$$q(\gamma+i\mu, t)=q(\gamma, t-2\mu)+2\int_0^\mu b(\gamma+i\mu', t-2(\mu-\mu'))\,d\mu' \tag{2.13}$$

for $\gamma+i\mu$ complex in which

$$g(\gamma, t)=\begin{cases}\xi(-t/2+i\gamma), & t>0,\\ -\xi(-t/2+i\gamma), & t<0,\end{cases} \tag{2.14}$$

$$\xi(z)=\frac{1}{2\pi}\frac{e^z}{1-e^z}. \tag{2.15}$$

If a and b are analytic in γ, then p and q defined by (2.10)-(2.13) are also analytic.

According to (2.5), $r=\frac{1}{2}(p+q)+i\frac{1}{2}(p^*-q^*)$. Let $A[s_0+r]$ denote the corresponding combination of the right-hand sides of (2.10)-(2.13). The Birkhoff-Rott equation (1.1) or (2.4) for $z=\gamma+s_0+r$ can be rewritten as follows:

$$r=A[s_0+r]. \tag{2.16}$$

Equations (2.10)-(2.13) for p, q or the equivalent equation (2.16) for r are the main results of this section.

3. Iteration method. In the previous section, the Birkhoff-Rott equation, for $z=\gamma+s_0+r$ was reduced to (2.16) for r. We now solve this equation by iteration. Define $r_0=0$ and for $n\geqq 0$ let r_{n+1} solve

$$r_{n+1}(\gamma, t)=A[s_0+r_n] \tag{3.1}$$

In terms of p_n, q_n, defined as in (2.5), equation (3.1) is written as in (2.10)-(2.13):

$$\begin{aligned} p_{n+1}(\gamma, t)&=\int_0^t\int_0^{2\pi} g(\gamma', t-t')a_n(\gamma-\gamma')\,d\gamma'\,dt',\\ q_{n+1}(\gamma, t)&=\int_t^\infty\int_0^{2\pi} g(\gamma', t-t')b_n(\gamma-\gamma')\,d\gamma'\,dt', \end{aligned} \tag{3.2}$$

$$\begin{aligned} p_{n+1}(\gamma+i\mu, t)&=p_{n+1}(\gamma, t+2\mu)-2\int_0^\mu a_n(\gamma+i\mu', t+2(\mu-\mu'))\,d\mu',\\ q_{n+1}(\gamma+i\mu, t)&=q_{n+1}(\gamma, t-2\mu)+2\int_0^\mu b_n(\gamma+i\mu', t-2(\mu-\mu'))\,d\mu' \end{aligned} \tag{3.3}$$

for γ and μ real, in which a_n, b_n are defined as in (2.8) with r replaced by r_n. To show convergence of r_n we obtain estimates on the difference

$$R_n=r_n-r_{n-1}.$$

Let $Q_n=q_n-q_{n-1}$, $P_n=p_n-p_{n-1}$. We use the following differentiated equations for R_n, or equivalently for P_n, Q_n:

$$\begin{aligned} \partial_\gamma^k P_{n+1}(\gamma, t)&=\int_0^t\int_0^{2\pi} g\partial_\gamma^k(a_n-a_{n-1})\,d\gamma'\,dt',\\ \partial_\gamma^k Q_{n+1}(\gamma, t)&=\int_t^\infty\int_0^{2\pi} g\partial_\gamma^k(b_n-b_{n-1})\,d\gamma'\,dt', \end{aligned} \tag{3.4}$$

$$\begin{aligned}\partial_\gamma^k P_{n+1}(\gamma+i\mu, t) &= \partial_\gamma^k P_{n+1}(\gamma, t+2\mu) - 2\int_0^\mu \partial_\gamma^k (a_n - a_{n-1})\, d\mu' \\ \partial_\gamma^k Q_{n+1}(\gamma+i\mu, t) &= \partial_\gamma^k Q_{n+1}(\gamma, t-2\mu) + 2\int_0^\mu \partial_\gamma^k (b_n - b_{n-1})\, d\mu'\end{aligned} \tag{3.5}$$

for $k = 1, 2$.

For the Hölder norm $|s|_\rho$ defined in (1.6), the Cauchy estimate for the derivative of an analytic function is

$$|s_\gamma(\cdot, t)|_\rho \leqq (\rho' - \rho)^{-1} |s(\cdot, t)|_{\rho'} \quad \text{if } \rho' > \rho. \tag{3.6}$$

The nonlinear part B_r of the Birkhoff-Rott integral operator is estimated as

$$\begin{aligned}&|B_r[s]|_\rho \leqq c_0 |s_\gamma|_\rho^2, \\ &|B_r[s]_\gamma|_\rho \leqq c_0 |s_\gamma|_\rho |s_{\gamma\gamma}|_\rho, \\ &|B_r[s] - B_r[\tilde{s}]|_\rho \leqq c_1 |s_\gamma - \tilde{s}_\gamma|_\rho (|s_\gamma|_\rho + |\tilde{s}_\gamma|_\rho), \\ &|B_r[s]_\gamma - B_r[\tilde{s}]_\gamma|_\rho \leqq c_1 |s_\gamma - \tilde{s}_\gamma|_\rho (|s_{\gamma\gamma}|_\rho + |\tilde{s}_{\gamma\gamma}|_\rho) + c_1 |s_{\gamma\gamma} - \tilde{s}_{\gamma\gamma}|_\rho (|s_\gamma|_\rho + |\tilde{s}_\gamma|_\rho)\end{aligned} \tag{3.7}$$

for any s, $\tilde{s}$ in which $c_0 = c(1 - |s_\gamma|_\rho)^{-1}$ and $c_1 = c(1 - |s_\gamma|_\rho)^{-1} + c(1 - |\tilde{s}_\gamma|_\rho)^{-1}$. For these estimates we assume that $|s_\gamma|_\rho < 1$, $|\tilde{s}_\gamma|_\rho < 1$.

From these general estimates it follows that a_n, b_n satisfy

$$\begin{aligned}&|a_0|_\rho + |b_0|_\rho \leqq c |s_{0\gamma}|_\rho^2, \\ &|a_{0\gamma}|_\rho + |b_{0\gamma}|_\rho \leqq c |s_{0\gamma}|_\rho |s_{0\gamma\gamma}|_\rho,\end{aligned} \tag{3.8}$$

$$|a_n - a_{n-1}|_\rho + |b_n - b_{n-1}|_\rho \leqq c |R_{n\gamma}|_\rho (|s_{0\gamma}|_\rho + |r_{n-1\gamma}|_\rho + |r_{n\gamma}|_\rho), \tag{3.9}$$

$$\begin{aligned}|(a_n - a_{n-1})_\gamma|_\rho + |(b_n - b_{n-1})_\gamma|_\rho \leqq{}& c |R_{n\gamma}|_\rho (|s_{0\gamma\gamma}|_\rho + |r_{n-1\gamma\gamma}|_\rho + |r_{n\gamma\gamma}|_\rho) \\ &+ c |R_{n\gamma\gamma}|_\rho (|s_{0\gamma}|_\rho + |r_{n-1\gamma}|_\rho + |r_{n\gamma}|_\rho)\end{aligned} \tag{3.10}$$

if $|r_{m\gamma}|_\rho < |s_{0\gamma}|_\rho < \frac{1}{2}$, $|r_{m\gamma\gamma}|_\rho < |s_{0\gamma\gamma}|_\rho < \frac{1}{2}$ for $m = n, n-1$.

For use in estimating the iterates r_n, we state a general lemma. Its proof is a straightforward extension of the proof of Hölder bounds on the Hilbert transform (Katznelson [7]).

LEMMA 3.1. *Suppose that* $\int_0^{2\pi} a\, d\gamma = \int_0^{2\pi} b\, d\gamma = 0$ *and let* p *and* q *satisfy*

$$\begin{aligned}p(\gamma, t) &= \int_0^t \int_0^{2\pi} g(\gamma', t-t') a(\gamma - \gamma', t')\, d\gamma'\, dt', \\ q(\gamma, t) &= \int_t^\infty \int_0^{2\pi} g(\gamma', t-t') b(\gamma - \gamma', t')\, d\gamma'\, dt'\end{aligned} \tag{3.11}$$

and suppose that

$$|a_\gamma|_0 + |b_\gamma|_0 \leqq c\varepsilon^2 (1 + t^{\nu-\alpha-1})\, e^{-t}. \tag{3.12}$$

Let r *be related to* p *and* q *as in* (2.5). *Then* r *satisfies*

$$|r_\gamma|_0 \leqq c\varepsilon^2 e^{-t/2}, \qquad |r_{\gamma\gamma}|_0 \leqq c\varepsilon^2 (1 + t^{\nu-\alpha-1})\, e^{-t/2}. \tag{3.13}$$

The same estimates are true for p *and* q.

4. First approximation. The first approximation $r_1 = R_1$ satisfies $r_1 = A[s_0]$, i.e.,

$$p_1(\gamma, t) = \int_0^t \int_0^{2\pi} g a_0 \, d\gamma' \, dt', \qquad q_1(\gamma, t) = \int_t^\infty \int_0^{2\pi} g b_0 \, d\gamma' \, dt', \tag{4.1}$$

$$\begin{aligned} p_1(\gamma + i\mu, t) &= p_1(\gamma, t+2\mu) - 2\int_0^\mu a_0 \, d\mu', \\ q_1(\gamma + i\mu, t) &= q_1(\gamma, t-2\mu) + 2\int_0^\mu b_0 \, d\mu' \end{aligned} \tag{4.2}$$

in which the arguments inside the integrals are as in (3.2)-(3.3). From (3.8), (1.7), and (1.8) it follows that a_0, b_0 satisfy

$$|a_{0\gamma}|_\rho + |b_{0\gamma}|_\rho \leqq c\varepsilon^2 (1 + (t-2\rho)^{\nu-\alpha-1}) \, e^{-(t-2\rho)},$$

for $t > 2\rho$. In particular,

$$|a_{0\gamma}|_0 + |b_{0\gamma}|_0 \leqq c\varepsilon^2 (1 + t^{\nu-\alpha-1}) \, e^{-t}.$$

Application of Lemma 3.1 implies

$$|r_{1\gamma}|_0 \leqq c\varepsilon^2 \, e^{-t/2} \qquad |r_{1\gamma\gamma}|_0 \leqq c\varepsilon^2 (1 + t^{\nu-\alpha-1}) \, e^{-t/2}. \tag{4.3}$$

Next estimate $r_{1\gamma}$ for $\gamma + i\mu$ complex from (4.2) as

$$\begin{aligned} |r_{1\gamma}(\cdot, t)|_\rho &\leqq 2 \sup_{|\mu| \leqq \rho} |r_{1\gamma}(\cdot, t+2\mu)|_0 + 2\int_0^\rho |a_{0\gamma}(t+2\mu')|_{\rho-\mu'} + |b_{0\gamma}(t-2\mu')|_{\rho-\mu'} \, d\mu' \\ &\leqq c\varepsilon^2 e^{-(t-2\rho)/2} + c\varepsilon^2 \rho (1 + (t-2\rho)^{\nu-\alpha-1}) \, e^{-(t-2\rho)} \\ &\leqq c\delta^{\nu-\alpha-1} \varepsilon^2 \, e^{-(t-2\rho)/2} \end{aligned} \tag{4.4}$$

for $\kappa_1 \rho < t$ with $\kappa_1 = 2(1+\delta)$ with $\delta > 0$. Similarly,

$$\begin{aligned} |r_{1\gamma\gamma}(\cdot, t)|_\rho &\leqq 2 \sup_{|\mu| \leqq \rho} |r_{1\gamma\gamma}(\cdot, t+2\mu)|_0 + 2\int_0^\rho |a_{0\gamma\gamma}(t+2\mu')|_{\rho-\mu'} + |b_{0\gamma\gamma}(t-2\mu')|_{\rho-\mu'} \, d\mu' \\ &\leqq c\varepsilon^2 (1 + (t-2\rho)^{\nu-\alpha-1}) \, e^{-(t-2\rho)/2} \\ &\quad + (\delta\rho)^{-1} \int_0^\rho (|a_{0\gamma}|_{(1+\delta)\rho-\mu'} + |b_{0\gamma}|_{(1+\delta)\rho-\mu'}) \, d\mu' \\ &\leqq c\varepsilon^2 (1 + (t-2\rho)^{\nu-\alpha-1}) \, e^{-(t-2\rho)/2} + 2\varepsilon^2 \delta^{-1} (1 + (t-\kappa_1\rho)^{\nu-\alpha-1}) \, e^{-(t-\kappa_1\rho)/2} \\ &\leqq c\delta^{-1} \varepsilon^2 (1 + (t-\kappa_1\rho)^{\nu-\alpha-1}) \, e^{-(t-\kappa_1\rho)/2} \end{aligned} \tag{4.5}$$

for $\kappa_1 \rho < t$.

5. Induction hypothesis. Successive approximations r_n and their convergence are analyzed using an analogue of the Cauchy-Kowalewski method [13] for the integral equations (3.2), (3.3). In each iteration the wedge of existence $|\text{Im } \gamma| < \kappa_n^{-1} t$ is slightly decreased, so that the general estimate (3.6) can be employed. The object of the modified Cauchy-Kowalewski method is to show that there is a limiting, nonempty wedge in which the solution r is analytic. The key point of the method is overcoming the large factor $(\rho - \rho')^{-1}$ in the Cauchy estimate (3.6). As a result of this factor the singularity in the solution r may be one order higher than the singularity in s_0 on the lines $|\text{Im } \gamma| = \kappa_n^{-1} t$. However, the additional singularity has amplitude proportional to t, so that at $t = 0$, r is no more singular than s_0.

The induction argument follows that of Nishida [13] with some change in notation and indexing. Define

$$\kappa_{n+1} = \kappa_n (1 - \varepsilon^{\mu}(n+2)^{-2})^{-1} \tag{5.1}$$

for $n \geqq 1$, in which $\mu > 0$ is to be chosen later. Then

$$\kappa_n \to \kappa = \kappa_1 \prod_1^{\infty} (1 - \varepsilon^{\mu}(m+2)^{-2})^{-1}, \tag{5.2}$$

which is finite and note that $\kappa_{n+1} > \kappa_n > \cdots > \kappa_1 > 2$. Define the norm

$$\begin{aligned} A_m(R) = \sup_{t > \kappa_m \rho} & \varepsilon^{-2} e^{(t-\kappa_m\rho)/2} t^{-1}(t - \kappa_m \rho) \\ & \times \{|R_\gamma(t)|_\rho + (1 + (t - \kappa_m\rho)^{\nu-\alpha-1})^{-1} |R_{\gamma\gamma}(t)|_\rho\}. \end{aligned} \tag{5.3}$$

Since $\kappa_{m+1} > \kappa_m$, then $A_{m+1}(R) \leqq A_m(R)$ for any R. The size of the mth correction $R_m = r_m - r_{m-1}$ is measured by

$$\lambda_m = A_m(R_m). \tag{5.4}$$

As mentioned above, the factor $(t - \kappa_m\rho)/t$ indicates possible additional singularity on the line $|\operatorname{Im} \gamma| = \rho = \kappa_m^{-1} t$, but note that it does not affect the size of R_m on the physical line $\operatorname{Im} \gamma = \rho = 0$.

The induction hypothesis is

$$\sum_{j=1}^{n} \lambda_j \left(1 - \frac{\kappa_j}{\kappa_{j+1}}\right) \leqq d. \tag{5.5}$$

The constant d as well as κ_1 will be chosen later. In the previous section, we showed that $\lambda_1 \leqq c(\kappa_1 - 2)^{-1}$ so that the hypothesis for $n = 1$ is satisfied if $d \geqq c(\kappa_1 - 2)^{-1}$.

Now suppose that (5.5) is true for $n \geqq 1$. If $\kappa_{n+1}\rho < t$, then for $n \geqq j \geqq 1$,

$$\begin{aligned} |R_{j\gamma}(t)|_\rho &< \lambda_j \left(1 - \frac{\kappa_j}{\kappa_{n+1}}\right)^{-1} \varepsilon^2 e^{-(t-\kappa_j\rho)/2} \\ &< \lambda_j \left(1 - \frac{\kappa_j}{\kappa_{j+1}}\right)^{-1} \varepsilon^2 e^{-(t-\kappa_n\rho)/2}, \end{aligned} \tag{5.6}$$

$$\begin{aligned} |R_{j\gamma\gamma}(t)|_\rho &< \lambda_j \left(1 - \frac{\kappa_j}{\kappa_{n+1}}\right)^{-1} \varepsilon^2 (1 + (t - \kappa_j\rho)^{\nu-\alpha-1}) e^{-(t-\kappa_j\rho)/2} \\ &< \lambda_j \left(1 - \frac{\kappa_j}{\kappa_{j+1}}\right)^{-1} \varepsilon^2 (1 + (t - \kappa_n\rho)^{\nu-\alpha-1}) e^{-(t-\kappa_n\rho)/2}. \end{aligned} \tag{5.7}$$

Since $r_0 = 0$, $r_{n\gamma} = \sum_{j=1}^{n} R_{j\gamma}$, $r_{n\gamma\gamma} = \sum_{j=1}^{n} R_{j\gamma\gamma}$, and using the induction hypothesis at n, this yields

$$\begin{aligned} |r_{j\gamma}(t)|_\rho &\leqq \varepsilon^2 \left(\sum_{j=1}^{n} \lambda_j (1 - \kappa_j/\kappa_{j+1})^{-1} \right) e^{-(t-\kappa_n\rho)/2} \\ &\leqq d\,\varepsilon^2 e^{-(t-\kappa_n\rho)/2}, \end{aligned} \tag{5.8}$$

$$|r_{j\gamma\gamma}(t)|_\rho \leqq d\varepsilon^2 (1 + (t - \kappa_n\rho)^{\nu-\alpha-1}) e^{-(t-\kappa_n\rho)/2}, \tag{5.9}$$

for any $j \leqq n$ and $\kappa_{n+1}\rho < t$. The bounds (5.8), (5.9) will be used in § 6 for estimation of R_{n+1}.

6. Successive approximations. To verify the induction hypothesis (5.5) for $n+1$, the correction terms R_{n+1} that solve (3.4), (3.5) must be estimated. Using (1.7), (1.8) for s_0, (5.8), (5.9) for r_n, r_{n-1}, and (5.4) for $R_{n\gamma}$, $R_{n\gamma\gamma}$ in (3.10), the forcing term in (3.4), (3.5) is bounded for $\kappa_{n+1}\rho < t$ by

$$|(a_n - a_{n-1})_\gamma|_\rho + |(b_n - b_{n-1})_\gamma|_\rho \tag{6.1}$$
$$\leqq c\varepsilon^3(1+\varepsilon d)\lambda_n\left(\frac{t}{t-\kappa_n\rho}\right)(1+(t-\kappa_n\rho)^{\nu-\alpha-1})\, e^{-(t-\kappa_n\rho)}.$$

In particular for $\rho = 0$,

$$|(a_n - a_{n-1})_\gamma|_0 + |(b_n - b_{n-1})_\gamma|_0 \leqq c\varepsilon^3(1+\varepsilon d)\lambda_n(1+t^{\nu-\alpha-1})\, e^{-t}. \tag{6.2}$$

First, estimate R_{n+1} for γ real, i.e., $\rho = 0$. Application of Lemma 3.1 to (3.4) using (6.2) implies that

$$|R_{n+1\gamma}|_0 \leqq c\varepsilon^3(1+\varepsilon d)\lambda_n\, e^{-t/2}, \tag{6.3}$$

$$|R_{n+1\gamma\gamma}|_0 \leqq c\varepsilon^3(1+\varepsilon d)\lambda_n(1+t^{\nu-\alpha-1})\, e^{-t/2}. \tag{6.4}$$

Second, estimate $R_{n+1\gamma}$ for complex $\gamma + i\mu$, i.e., for $\rho \geqq 0$, solving (3.5). For $\kappa_{n+1}\rho < t$ bound

$$|R_{n+1\gamma}(t)|_\rho \leqq 2\sup_{|\mu|<\rho}|R_{n+1\gamma}(t+2\mu)|_0 + \int_0^\rho (|(a_n - a_{n-1})_\gamma(t+2\mu)|_{\rho-\mu}$$
$$+ |(b_n - b_{n-1})_\gamma(t-2\mu)|_{\rho-\mu})\, d\mu \tag{6.5}$$
$$\leqq 2c\varepsilon^3(1+\varepsilon d)\lambda_n\, e^{-(t-\rho)/2} + I,$$

in which I denotes the integral. The inequalities $\kappa_{n+1} > \kappa_n > 2$ and $\kappa_{n+1}\rho < t$ imply $\kappa_{n+1}(\rho-\mu) < (t-2\mu)$. Then we may use (6.1) to bound I by

$$I \leqq c\varepsilon^3(1+\varepsilon d)\lambda_n \int_0^\rho \left(\frac{t\pm 2\mu}{(t\pm 2\mu) - \kappa_n(\rho-\mu)}\right)(1+(t\pm 2\mu - \kappa_n(\rho-\mu))^{\nu-\alpha-1})$$
$$\cdot\, e^{-(t\pm 2\mu - \kappa_n(\rho-\mu))}\, d\mu$$
$$\leqq c\varepsilon^3(1+\varepsilon d)\lambda_n(t+\rho)\, e^{-(3/4)(t-\kappa_n\rho)} \int_0^\rho (t-\kappa_n\rho + (\kappa_n - 2)\mu)^{\nu-\alpha-2}\, d\mu \tag{6.6}$$
$$\leqq c\varepsilon^3(1+\varepsilon d)(\kappa_n-2)^{-1}\lambda_n(t+\rho)(t-\kappa_n\rho)^{\nu-\alpha-1}\, e^{-(3/4)(t-\kappa_n\rho)}$$
$$\leqq c\varepsilon^3(1+\varepsilon d)(\kappa_n-2)^{-1}\lambda_n\left(\frac{t}{t-\kappa_n\rho}\right) e^{-(t-\kappa_n\rho)/2}.$$

Since $\kappa_{n+1} > \kappa_n$, this combines with (6.5) to show that

$$|R_{n+1\gamma}(t)|_\rho \leqq c\varepsilon^3(1+\varepsilon d)(\kappa_n-2)^{-1}\lambda_n\left(\frac{t}{t-\kappa_{n+1}\rho}\right) e^{-(t-\kappa_{n+1}\rho)/2}, \tag{6.7}$$

for $\kappa_{n+1}\rho < t$.

Third, estimate $R_{n+1\gamma\gamma}$ for complex $\gamma + i\mu$, i.e., for $\rho \geqq 0$, solving (3.5). This is the crucial estimate of the Cauchy-Kowalewski method. For $\kappa_{n+1}\rho < t$ estimate

$$|R_{n+1\gamma\gamma}(t)|_\rho \leqq 2\sup_{|\mu|\leqq\rho}|R_{n+1\gamma\gamma}(t+2\mu)|_0$$
$$+ \int_0^\rho \{|(a_n - a_{n-1})_{\gamma\gamma}(t+2\mu)|_{\rho-\mu} + |(b_n - b_{n-1})_{\gamma\gamma}(t-2\mu)|_{\rho-\mu}\}\, d\mu \tag{6.8}$$
$$\leqq c\varepsilon^3(1+\varepsilon d)\lambda_n(1+(t-2\rho)^{\nu-\alpha-1})\, e^{-(t-2\rho)/2} + I_2,$$

in which I_2 is the integral. Define ρ_1 by

$$\rho_1 = \tfrac{1}{2}(\kappa_{n+1}^{-1}(t-2\mu)+(\rho-\mu)), \tag{6.9}$$

which satisfies

$$\rho-\mu < \rho_1 < \kappa_{n+1}^{-1}(t-2\mu), \tag{6.10}$$

$$\begin{aligned} t-2\mu-\kappa_{n+1}\rho_1 &= \kappa_{n+1}(\rho_1-(\rho-\mu)) = (t-\kappa_{n+1}\rho+(\kappa_{n+1}-2)\mu)/2 \\ &\geqq (t-\kappa_{n+1}\rho)/2. \end{aligned} \tag{6.11}$$

Estimate I_2 using the Cauchy estimate (3.6), the bound (6.1), and the relations (6.10), (6.11) to obtain

$$\begin{aligned} I_2 &\leqq 2\int_0^\rho (\rho_1-(\rho-\mu))^{-1}|(a_n-a_{n-1})_\gamma(t+2\mu)|_{\rho_1}+|(b_n-b_{n-1})_\gamma(t-2\mu)|_{\rho_1}\,d\mu \\ &\leqq c\varepsilon^3(1+\varepsilon d)\lambda_n\int_0^\rho (\rho_1-(\rho-\mu))^{-1}\frac{t+2\mu}{(t-2\mu-\kappa_n\rho_1)}(1+(t-2\mu-\kappa_n\rho_1)^{\nu-\alpha-1}) \\ &\quad\cdot e^{-(t-2\mu-\kappa_{n+1}\rho_1)}\,d\mu \\ &\leqq c\varepsilon^3(1+\varepsilon d)\lambda_n(t+\rho)\,e^{-(t-\kappa_{n+1}\rho)/2}\int_0^\rho (t-\kappa_{n+1}\rho+(\kappa_{n+1}-2)\mu)^{-2} \\ &\quad+(t-\kappa_{n+1}\rho+(\kappa_{n+1}-2)\mu)^{\nu-\alpha-3}\,d\mu \\ &\leqq c\varepsilon^3(1+\varepsilon d)\lambda_n(t+\rho)\,e^{-(t-\kappa_{n+1}\rho)/2}(\kappa_{n+1}-2)^{-1} \\ &\quad\cdot(t-\kappa_{n+1}\rho)^{-1}(1+(t-\kappa_{n+1}\rho)^{\nu-\alpha-1}). \end{aligned} \tag{6.12}$$

Combine this with (6.8) to find

$$\begin{aligned} |R_{n+1\gamma\gamma}|_\rho &\leqq c\varepsilon^3(1+d)\lambda_n(\kappa_{n+1}-2)^{-1}\left(\frac{t}{t-\kappa_{n+1}\rho}\right) \\ &\quad\cdot(1+(t-\kappa_{n+1}\rho)^{\nu-\alpha-1})\,e^{-(t-\kappa_{n+1}\rho)/2} \end{aligned} \tag{6.13}$$

for $\kappa_{n+1}\rho < t$.

Inequalities (6.7) for $R_{n+1\gamma}$ and (6.13) for $R_{n+1\gamma\gamma}$ and the definition (5.4) of λ_{n+1} imply that

$$\begin{aligned} \lambda_{n+1} &\leqq c\varepsilon(1+\varepsilon d)(\kappa_{n+1}-2)^{-1}\lambda_n \\ &\leqq c\varepsilon(1+\varepsilon d)(\kappa_1-2)^{-1}\lambda_n. \end{aligned} \tag{6.14}$$

This inequality is also true for λ_{j+1} in terms of λ_j for any $j \leqq n$, so that

$$\begin{aligned} \lambda_{n+1} &\leqq \{c\varepsilon(1+\varepsilon d)(\kappa_1-2)^{-1}\}^n\lambda_1 \\ &\leqq \{c\varepsilon(1+\varepsilon d)(\kappa_1-2)^{-1}\}^n(\kappa_1-2)^{-1}. \end{aligned} \tag{6.15}$$

With these estimates finished, we are ready to verify the induction hypothesis (5.5) for $n+1$ by choosing d and κ_1. Let

$$\kappa_1 = 2+a\varepsilon^\mu, \qquad d = \varepsilon^{-2\mu} \tag{6.16}$$

in which a and the parameters μ from (5.1) are still to be chosen.

Estimate

$$\sum_{j=1}^{n+1} \lambda_j(1-\kappa_j/\kappa_{j+1})^{-1} \leqq (\kappa_1-2)^{-1} \sum_{j=1}^{\infty} (c\varepsilon(1+\varepsilon d)(\kappa_1-2)^{-1})^{j-1}(1-\kappa_j/\kappa_{j+1})^{-1}$$

$$= a^{-1}\varepsilon^{-2\mu} \sum_{j=1}^{\infty} (ca^{-1}\varepsilon^{1-\mu}(1+\varepsilon^{1-2\mu}))^{j-1}(j+2)^2 \tag{6.17}$$

$$< \varepsilon^{-2\mu}$$

for any μ with $0 \leqq \mu \leqq 2/3$ and a chosen (independently of μ) to be sufficiently large. This verifies the induction hypothesis (5.5) for $(n+1)$, for any n.

Completion of the induction proof shows that the inequalities (5.6), (5.7) on the jth corrections R_j are valid for all j and that therefore the approximate solutions r_n have a limit r that solves (2.16). Finally the combination $z(\gamma, t) = \gamma + s_0(\gamma, t) + r(\gamma, t)$ solves the Birkhoff-Rott equation (1.1).

Moreover, the solution $r = \lim_{n\to\infty} r_n$ is analytic in $\kappa\rho < t$ with

$$\kappa = \lim_{m\to\infty} \kappa_m = \kappa_1 \Pi(1-\varepsilon^{\mu}(m+2)^2)^{-1}$$

$$= 2 + Q(\varepsilon^{\mu}).$$

The bounds (5.8), (5.9) show that for $\kappa\rho < t$, r satisfies

$$|r(t)|_{\rho} + |r_{\gamma}(t)|_{\rho} \leqq c\varepsilon^{2-2\mu}\, e^{-(t-\kappa\rho)/2}$$

$$|r_{\gamma\gamma}(t)|_{\rho} \leqq c\varepsilon^{2-2\mu}(1+(t-\kappa\rho)^{\nu-\alpha-1})\, e^{-(t-\kappa\rho)/2}. \tag{6.18}$$

By choosing $\mu = \frac{2}{3}$, its largest permissible size, we find that r is analytic in the region $(2+Q(\varepsilon^{2/3}))\ \rho < t$. By choosing $\mu = 0$, we obtain the optimal bounds on the size of r, although on a restricted domain. In particular for $\rho = 0$, i.e., on the physical line γ real, the bounds are

$$|r(t)|_0 + |r_{\gamma}(t)|_0 \leqq c\varepsilon^2\, e^{-t/2}$$

$$|r_{\gamma\gamma}(t)|_0 \leqq c\varepsilon^2(1+t^{\nu-\alpha-1})\, e^{-t/2}. \tag{6.19}$$

This completes the proof of Theorem 1.

7. Conclusions. We will use Theorem 1 to derive solutions of the Birkhoff-Rott equation that develop singularities (i.e., infinite curvature) at finite time starting from analytic initial data. Then we show that the initial value problem for this equation is ill posed in Sobolev space H^n $(n > \frac{3}{2})$, since the derivative of order $1+\nu$, for any $\nu > 0$, can become infinite in an arbitrarily small time from arbitrarily small initial data (fractional derivatives are understood in the Hölder sense). On the other hand, the vortex sheet problem is known to be well posed in an analytic function setting for at least a short time [2], [14]. DiPerna and Majda [3]-[5] address the related question of finding an appropriate function space that is preserved by the Euler flow and by limits of Euler solutions, as well as by limits of regularized solutions (i.e., of the Navier-Stokes equations or the numerical vortex method).

The Birkhoff-Rott equation has the following three symmetry properties: If $z(\gamma, t)$ is a solution of (1.1) then so are $z_b(\gamma, t) = z^*(\gamma, -t)$, $z_s(\gamma, t) = z(\gamma, t-t_0)$, and $z_n(\gamma, t) = n^{-1}z(n\gamma, nt)$.

When we use z_b, Corollary 1 follows from Theorem 1.

COROLLARY 1. *Let ε, ν, α be as in Theorem* 1 *and let s_0 satisfy properties* (i)-(iii) *except for $t < 0$; i.e., s_0 is an analytic solution of the linearized equation* (1.3) *which decays to zero at $t = -\infty$ (decaying backwards in time) and has a mild singularity at $t = 0$. Then*

there is a function $r(\gamma, t)$ *such that* $z(\gamma, t) = \gamma + s_0 + r$ *is an analytic solution of the Birkhoff-Rott equation* (1.1) *for* $t < 0$ *and* $\kappa |\operatorname{Im} \gamma| < |t|$ *in which* $\kappa > 2$ *and* $\kappa \to 2$ *as* $\varepsilon \to 0$. *Moreover,* r *can be chosen so that the backward decaying mode* $r_+ - i\bar{r}_- = 0$ *at* $t = 0$ *and so that* r *satisfies* (1.10), (1.11) *for* $t < 0$.

By shifting the origin of time in the solution z of Corollary 1, we obtain Corollary 2.

COROLLARY 2. *There is initial data* $z(\gamma, 0)$, *which is analytic in a neighborhood of* γ *real, such that the solution* $z(\gamma, t)$ *of the Birkhoff-Rott equation* (1.1) *develops an infinite* $(1+\nu)$*th derivative at a finite time* t_0.

Finally we use the rescaling of z to z_n. Take z to be a solution for $t < 0$ that develops an infinite $(1+\nu)$th derivative at $t = 0$, as in Corollary 1. Let $z_N(\gamma, t) = N^{-2} z(N^2\gamma, N^2 t - 2N)$ so that $s_N = z_N - \gamma = N^{-2} s(N^2\gamma, N^2 t - 2N)$. Then at $t = 0$ the kth Sobolev norm of s_N is bounded as

$$
\begin{aligned}
|s_N(\cdot, t=0)|_{H^k} &= N^{-2+2k+3} |s(\cdot, -2N)|_{H^k} \\
&\leqq N^{2k+1} e^{-N} \qquad (7.1) \\
&\to 0 \quad \text{as } N \to \infty.
\end{aligned}
$$

However, the time T_N of singularity formation is $T_N = 2N^{-1} \to 0$ as $N \to \infty$. For $\nu > 0$, denote $\sup |\partial_\gamma^{1+\nu} z| = \sup_{\gamma \neq \gamma' \text{ real}} |z'(\gamma) - z'(\gamma')| / |\gamma - \gamma'|^\nu$. This shows the following.

COROLLARY 3. *For any positive numbers* ν, k, ε, *and* δ *there is initial data* $z = \gamma + s_0$ *with* $|s|_{H^k} < \varepsilon$ *such that* $\sup |\partial_\gamma^{1+\nu} z|$ *goes to infinity for* $t = t_0 < \delta$. *In particular the initial value problem for the Birkhoff-Rott equation* (1.1) *is ill posed for any Sobolev space* H^k *for* $k > 3/2$.

In other words, smallness of the initial perturbation s is not sufficient to insure existence with bounded $(1+\nu)$th derivative on any time interval for (1.1).

REFERENCES

[1] G. BIRKHOFF, *Helmholtz and Taylor instability* in Hydrodynamic Instability, Proc. Sympos. in Appl. Math. XII, American Mathematical Society, Providence, RI, 1962, pp. 55-76.

[2] R. CAFLISCH AND O. ORELLANA, *Long time existence for a slightly perturbed vortex sheet*, Comm. Pure Appl. Math., 39 (1986), pp. 807-838.

[3] R. DIPERNA AND A. MAJDA, *Oscillations and concentrations in weak solutions of the incompressible fluid equations*, Comm. Math. Phys., 108 (1987), pp. 667-689.

[4] ———, *Concentrations and regularization for 2-D incompressible flow*, Comm. Pure Appl. Math., 40 (1987), pp. 301-345.

[5] ———, *Reduced Hausdorff dimension and concentration cancellation for two-dimensional incompressible flow*, J. Amer. Math. Soc., 1 (1988), pp. 59-86.

[6] J. DUCHON AND R. ROBERT, *Solution globales avec nappe tourbillionaire pour les equations d'Euler dans le plan*, C.R. Acad. Sci. Paris, 302 (1986), pp. 183-186.

[7] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, Dover Press, New York, 1968.

[8] R. KRASNY, *On singularity formation in a vortex sheet and the point vortex approximation*, J. Fluid Mech., 167 (1986), pp. 65-93.

[9] ———, *Desingularization of periodic vortex sheet roll-up*, J. Comp. Phys., 65 (1986), pp. 292-313.

[10] D. I. MEIRON, G. R. BAKER, AND S. A. ORSZAG, *Analytic structure of vortex sheet dynamics, Part 1, Kelvin-Helmholtz instability*, J. Fluid Mech., 114 (1982), pp. 283-298.

[11] D. W. MOORE, *The spontaneous appearance of a singularity in the shape of an evolving vortex sheet*, Proc. Roy. Soc. London Ser. A, 365 (1979), pp. 105-119.

[12] ———, *Numerical and analytical aspects of Helmholtz instability*, in Theoretical and Applied Mechanics, Proc. XVI Internat. Congr. Theoret. Appl. Mech., F. I. Niordson and N. Olhoff, eds., North-Holland, Amsterdam, 1984, pp. 629-633.

[13] T. NISHIDA, *A note on a theorem of Nirenberg*, J. Differential Geometry, 12 (1977), pp. 629-633.

[14] C. SULEM, P. L. SULEM, C. BARDOS, AND U. FRISCH, *Finite time analyticity for the two and three dimensional Kelvin-Helmholtz instability*, Comm. Math. Phys., 80 (1981), pp. 485-516.

[15] D. EBIN, *Ill-posedness of the Rayleigh-Taylor and Helmholtz problems for incompressible fluids* 13 (1988), pp. 1265-1295.

[16] J. DUCHON AND R. ROBERT, *Global vortex solutions of Euler equations in the plane*, Comm. Partial Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 308-353, March 1989

003

ARTIFICIAL BOUNDARY CONDITIONS FOR INCOMPRESSIBLE VISCOUS FLOWS*

LAURENCE HALPERN† AND MICHELLE SCHATZMAN‡

Abstract. Artificial boundary conditions for the linearized incompressible Navier-Stokes equations are designed by approximating the symbol of the transparent operator. The related initial boundary value problems are well posed in the same spaces as the original Cauchy problem. Furthermore, error estimates for small viscosity are proved.

Key words. incompressible Navier-Stokes equation, artificial boundary conditions, initial boundary value problems, Oseen approximation

AMS(MOS) subject classifications. 35Q10, 35G15, 76D99

1. Introduction.

Many problems arising in fluid mechanics lead to the resolution of a partial differential equation in an unbounded domain. Depending on the applications, various strategies have been developed.

For stationary flows around a body, integral equations are often used (see for instance [6]) or we can also bound the domain and prescribe on the artificial boundary the so-called "transparent" boundary condition, i.e., the boundary condition which simulates the missing part of the domain. This boundary condition is integral on the boundary, and the associated initial boundary value problem is solved numerically using either an eigenvalues expansion on the boundary [2], [22], [10] or a coupling between finite elements in the interior and integral equations on the boundary [18], [26].

For time dependent problems, the "transparent" boundary condition is integral in time and space and thus impractical in general. Tremendous research effort has attempted to design useful boundary conditions for inviscid flows. Most of the studies rely on a linearization of the equation near the boundary (for a nonlinear treatment of the problem see [16] and [29]). Of course in the applications we solve the nonlinear equations in the interior together with the linear boundary conditions.

There are two mathematical frames for these studies. On one hand, Engquist and Majda in [8] and [9] designed absorbing boundary conditions with wave propagation tools. On the other hand, Bayliss and Turkel in [4] and [5] used far field expansions. Both works write sequences of boundary conditions that are local (i.e., differential) in time and space on the boundary. This feature is due to the hyperbolicity or quasi-hyperbolicity of the operators they handle.

The problem becomes less clear when it comes to viscous flows. Some numerical answers have been given for compressible fluids (see Rudy and Strikwerda [25]). On the other hand, calculations have been performed in [11], [12], and [19] on the case of a parabolic equation. Moreover, the case of linear advection-diffusion has been treated in [13]. For incompressible flows it is still, as far as we know, an open mathematical question.

Our previous results were announced in [15].

We are concerned here with the incompressible Navier-Stokes equation,

* Received by the editors July 27, 1987; accepted for publication June 1, 1988.

† Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau Cedex, France.

‡ Centre de Mathématiques Appliquées et Industrielles, Université Claude Bernard, 69622 Villeurbanne Cedex, France.

for $N=2$ or 3:

$$\text{(1.1)}\qquad \begin{cases} u_t+(u\cdot\nabla)u-\nu\Delta u+\nabla p=f & \text{in } \mathbb{R}^N\times]0,\ T[,\\ \operatorname{div}(u)=0 & \text{in } \mathbb{R}^N\times]0,\ T[,\\ u(0)=u^0 & \text{in } \mathbb{R}^N, \end{cases}$$

where $u=(u_1,\cdots,u_N)$.

It is important to set the artificial boundaries outside of turbulent regime, sufficiently far for the flow to be considered as constant. We are then allowed to linearize around the constant state and to consider the Oseen approximation as follows.

$$\text{(1.2)}\qquad \begin{cases} u_t+(a\cdot\nabla)u-\nu\Delta u+\nabla p=f & \text{in } \mathbb{R}^N\times]0,\ T[,\\ \operatorname{div}(u)=0 & \text{in } \mathbb{R}^N\times]0,\ T[,\\ u(0)=u^0 & \text{in } \mathbb{R}^N. \end{cases}$$

The data f and u^0 are supposed to be compactly supported.

All throughout this paper we study the model problem: writing artificial boundary conditions on the hyperplane $\mathbb{R}^{N-1}$, i.e., such that the Oseen equation in the half-space $\mathbb{R}^N_-=\{(x,y)\in\mathbb{R}^N, x<0\}$ with the boundary condition prescribed on the hyperplane $x=0$ is an approximation of the Oseen equation in $\mathbb{R}^N$. This enables us to use the Fourier transform as an essential tool. The problem of designing artificial conditions on a closed artificial boundary will be treated in a forthcoming paper.

In [13], as a first step, the linear advection diffusion equation has been studied. This article concluded with a family of approximate boundary conditions that are local in time and space. In contrast, here the divergence-free condition implies a coupling, which makes the analysis more troublesome. In particular, the symbol of the operator to be approximated contains $|k|$, with k the dual variable of the tangential spatial variable, and it does not seem easy to approximate this symbol by polynomials or rational fractions of low degree in space. Thus, the approximate boundary conditions will be local in time and global in the tangential space variables.

In the course of justifying our calculations precisely, we will prove a number of interesting results on the spaces of divergence-free functions on a half space, without a Dirichlet boundary condition.

The analysis is made in $\mathbb{R}^2$, but is valid in $\mathbb{R}^3$ with slight modifications (see [14]).

In § 2, we define the spaces of divergence-free functions in $\mathbb{R}^2$ and prove trace theorems on $\Gamma=\{(x,y), x=0\}$. We then introduce the Oseen equation, give the formalism necessary for a variational formulation, and state a well-posedness theorem for the Cauchy problem in $\mathbb{R}^2$. We finish by specific trace results for the solution of Stokes equation, which emphasize the regularity of $u_1+\mathcal{H}u_2$ ($\mathcal{H}$ is the Hilbert transform along the boundary), solution of a heat equation.

In § 3 we compute the transparent boundary condition on Γ according to the following principle: Problem (1.2) in $\mathbb{R}^2$ is equivalent (in a sense we will make precise) to the transmission problem in $\Omega_-\times\Omega_+$, where $\Omega_\pm=\{(x,y), \pm x>0\}$ and

$$\begin{aligned} &\frac{\partial u_-}{\partial t}+(a\cdot\nabla)u-\nu\Delta u_-+\nabla p_-=f && \text{in } \Omega_-\times]0,\ T[,\\ &\operatorname{div}(u_-)=0 && \text{in } \Omega_-\times]0,\ T[,\\ &\frac{\partial u_+}{\partial t}+(a\cdot\nabla)u_+-\nu\Delta u_++\nabla p_+=0 && \text{in } \Omega_+\times]0,\ T[,\\ &\operatorname{div}(u_+)=0 && \text{in } \Omega_+\times]0,\ T[, \end{aligned}$$

with the initial data

$$\begin{cases} u_-(0) = u^0 & \text{in } \Omega_-, \\ u_+(0) = 0 & \text{in } \Omega_+, \end{cases}$$

and the transmission conditions

$$\begin{cases} u_-|_\Gamma = u_+|_\Gamma, \\ \sigma_n(u_-)|_\Gamma = \sigma_n(u_+)|_\Gamma, \end{cases}$$

where σ_n is the normal constraint.

We study the Oseen equation for u_+ in Ω_+, with nonhomogeneous Dirichlet boundary conditions, and prove the well-posedness in spaces where the partial Fourier-Laplace transform in time and tangential space is permissible. We write then a pseudodifferential relation between u_+ and $\sigma_n(u_+)$ on Γ. Thanks to the transmission conditions, it leads to the same pseudo-differential relation between u_- and $\sigma_n(u_-)$. We call it the transparent boundary condition after proving the uniqueness for the related initial boundary value problem in Ω_-.

In § 4 we design a family of approximations to the transparent boundary condition by approximating its symbol. These approximations are local in time and integral along the boundary. Using the tools developed in the previous sections we prove them to be well-posed with the same regularity as the solution of the Cauchy problem in $\mathbb{R}^2$. Even at low order these boundary conditions appear to be good approximations for small viscosity.

Finally in Appendix A, we give some information on Beppo-Levi spaces, and in Appendix B, we show why we cannot use spaces of fast-decreasing functions at infinity in the analysis of the present incompressible problems.

2. Definitions, notations and basic results. In this section, we describe a number of functional spaces which are useful for our study of Stokes problem with an advection term, otherwise called an Oseen system. We need spaces in which we are able to treat nonhomogeneous boundary conditions, and thus transparent and artificial boundary conditions. The typical space is the space of divergence-free functions, which are square integrable, with a square integrable gradient, but without any boundary condition. We give density results relative to these spaces, and corresponding trace results. Then we give rather classical results on the solution of the Oseen problem, in $\mathbb{R}^2$. Finally, if u is the solution of this problem, and if $\mathcal{H}$ is the Hilbert transform in the direction x_2, the function $u_1 + \mathcal{H}u_2$ has a number of special properties which will be useful everywhere in the sequel. In particular, if the support of the data of the Oseen problem is in the region $\{x_1 \leqq -X < 0\}$, the restriction of $u_1 + \mathcal{H}u_2$ to the boundary $\{x_1 = 0\} \times \mathbb{R} \times \mathbb{R}$ is arbitrarily smooth.

2.1. Classical functional spaces. We use the formalism of [24] in many instances. The domain W of $\mathbb{R}^N$ has boundary Γ; the scalar product in $L^2(\Omega)$ is denoted $(\cdot, \cdot)$, with associated norm $\| \ \|$. The classical Sobolev space $H^m(\Omega)$ is the space of square integrable functions, whose derivatives of order at most m are square integrable; the scalar product in $H^m(\Omega)$ is denoted $(\cdot, \cdot)_m$, and the norm $\| \ \|_m$; the scalar product in $L^2(\Omega)$ is denoted $(\cdot, \cdot)_\Omega$ with associated norm $\| \ \|_\Omega$. The Fourier transform is defined on the Schwarz space $\mathcal{S}$ by

$$\hat{v}(k) = \int_{\mathbb{R}^N} v(x) \exp(-ik \cdot x)\, dx, \tag{2.1}$$

where $k \cdot x = k_1 x_1 + k_2 x_2 + \cdots + k_N x_N$; it is extended to $\mathcal{S}'$, and we will often write (2.1) for temperate distributions by an abuse of notation.

For any real s, the Sobolev space of fractional order $H^s(\mathbb{R}^N)$ is defined as

$$H^s(\mathbb{R}^N) = \left\{ v \in \mathcal{S}' \Big/ \int (1+|k|^2)^s |\hat{v}(k)|^2 \, dk < \infty \right\},$$

this space is a Hilbert space, equipped with the norm

$$\|v\|_s = \left[\int (1+|k|^2)^s |\hat{v}(k)|^2 \, dk \right]^{1/2}.$$

In particular, we will need the Sobolev spaces $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$, when Γ is $\mathbb{R}^N$; the norm of $H^{1/2}(\Gamma)$ will be denoted $\|\cdot\|_{1/2,\Gamma}$, and the norm of $H^{-1/2}(\Gamma)\|\cdot\|_{-1/2,\Gamma}$.

We denote the right-hand side half-space as

$$\Omega_+ = \mathbb{R}^N_+ = \{x = (x_1, x_2, \cdots, x_N)/x_1 > 0\},$$

and the left-hand side half-space as

$$\Omega_- = \mathbb{R}^N_- = \{x = (x_1, x_2, \cdots, x_N)/x_1 < 0\}.$$

A convenient characterization of $H^m(\Omega_+)$ is given by

$$H^m(\Omega_+) = \{u \in L^2(0, \infty; H^m(\mathbb{R}^{N-1}))/ \quad \forall k = 1, 2, \cdots, m,$$

$$\frac{\partial^k u}{\partial x_1^k} \in L^2(0, \infty; H^{m-k}(\mathbb{R}^{N-1}))\}.$$

Then the norm on $H^m(\Omega_+)$ can be written as

$$\|u\|_m^2 = \sum \left| \frac{\partial^k u}{\partial x_1^k} \right|^2_{L^2(0,\infty;\, H^{m-1}(\mathbb{R}^{N-1}))}.$$

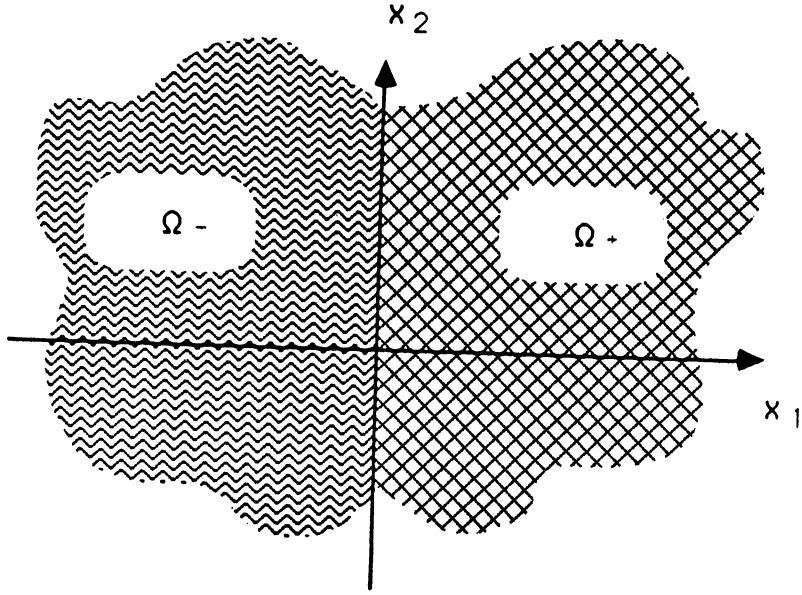


FIG. 1

We can define a partial Fourier transform operating on the tangential variables; for v in $\mathcal{S}'(\Omega_+)$, it is defined formally as

$$\mathcal{F}_{x'\to k'}v(x_1, k') = \int_{\mathbb{R}^{N-1}} v(x_1, x') \exp(-ik\cdot x')\, dx',$$

where $x' = (x_2, \cdots, x_N)$ and $k' = (k_2, \cdots, k_N)$.
The space $H^m(\Omega_+)$ can be characterized too with the help of a partial Fourier transform.

We define the Sobolev spaces of vector-valued functions:

$$\mathbf{H}^s(\Omega) = (H^s(\Omega))^N \text{ and } \mathbf{H}_0^s(\Omega) = (H_0^s(\Omega))^N.$$

The scalar product and the associated norm will be denoted as for the scalar Sobolev spaces.

We denote the gradient by ∇, and the divergence by $\nabla\cdot$.

Finally, we will need the Beppo-Levi spaces, defined as follows.

DEFINITION 2.1. Let Ω be an open domain of $\mathbb{R}^N$. The Beppo-Levi space $BL(H^s(\Omega))$ constructed on $H^s(\Omega)$ is the space of distributions u such that ∇u belongs to $H^s(\Omega)$.

The space $BL(H^s(\Omega))$ is a subspace of $H^{s+1}_{loc}(\Omega)\cap\mathcal{S}'(\Omega)$; more information on the Beppo-Levi spaces is given in Appendix A. Let it suffice here to observe that $BL(H^s(\Omega))$ is a Hilbert space, with Hilbert seminorm $\|\nabla u\|_s$; the kernel of this seminorm is the space of constant functions. Moreover, if Ω is unbounded, $BL(H^s(\Omega))$ contains unbounded elements. Examples of such behavior are given in Appendix A.

2.2. Spaces of divergence-free functions. Spaces of divergence-free functions are an essential tool for the study of fluid motion in Eulerian coordinates. One classical set of spaces is defined starting from

$$\mathcal{V}(\Omega) = \{v\in\mathcal{D}(\Omega)^2/\nabla\cdot v = 0\}.$$

The respective closures of v in $\mathbf{L}^2(\Omega)$ and $\mathbf{H}^1(\Omega)$ are denoted $H_0(\Omega)$ and $V(\Omega)$; they are very well adapted to the study of a problem with Dirichlet boundary conditions; they admit the following characterization, when $\Omega = \mathbb{R}^2$.

$$\begin{aligned} V(\mathbb{R}^2) &= \{v\in\mathbf{H}^1(\mathbb{R}^2)/\nabla\cdot v = 0\} \\ H_0(\mathbb{R}^2) &= \{v\in\mathbf{L}^2(\mathbb{R}^2)/\nabla\cdot v = 0\} \end{aligned} \tag{2.2}$$

When Ω is a domain with smooth boundary Γ, the characterization is modified as follows.

$$\begin{aligned} V(\Omega) &= \{v\in\mathbf{H}_0^1(\Omega)/\nabla\cdot v = 0\} \\ H_0(\Omega) &= \{v\in\mathbf{L}^2(\Omega)/\nabla\cdot v = 0 \text{ and } v\cdot n = 0 \text{ on } \Gamma\}, \end{aligned} \tag{2.3}$$

where n is the exterior normal to Γ.
This last characterization makes sense because, on the space

$$E(\Omega) = \{v\in\mathbf{L}^2(\Omega)/\nabla\cdot v\in L^2(\Omega)\},$$

the normal trace $v\cdot n$ is defined, belongs to $\mathbf{H}^{1/2}(\Gamma)$, and the mapping

$$v\to v\cdot n$$

from $E(\Omega)$ to $\mathbf{H}^{1/2}(\Gamma)$ is onto.

A convenient reference for the above classical results is [28]. Nevertheless, we are interested in boundary conditions which are not Dirichlet boundary conditions, and we introduce spaces which are not classical:

$$\mathcal{W}(\Omega_-) = \{v\in\mathcal{D}(\Omega_-)^2/\nabla\cdot v = 0\}, \tag{2.4}$$

where Ω_- is the closed half-plane. Moreover, we introduce

$$W(\Omega_-) = \{v \in \mathbf{H}^1(\Omega_-)/\nabla \cdot v = 0\} \tag{2.5}$$

$$H(\Omega_-) = \{v \in \mathbf{L}^2(\Omega_-)/\nabla \cdot v = 0\}. \tag{2.6}$$

The relations between $\mathcal{W}(\Omega_-)$, $W(\Omega_-)$, and $H(\Omega_-)$ as well as the trace properties of these spaces will be stated below. The results presented here differ from the standard ones in two respects: first, we want to allow for nonzero boundary data; second, the open set Ω contains points at an infinite distance from its boundary. The case with zero boundary data can be found in [28]; the case where the points of Ω are at a bounded distance from its boundary can be found in [1].

PROPOSITION 2.2. *The closure of $\mathcal{W}(\Omega_-)$ in $\mathbf{H}^1(\Omega_-)$ is precisely $W(\Omega_-)$.*

Proof. To prove this result, we consider an element orthogonal to the closure of $\mathcal{W}(\Omega_-)$ in $W(\Omega_-)$. The potential of this element u satisfies a homogeneous partial differential equation; as u is in a Sobolev space, it is a temperate distribution. This enables us to deduce estimates on the traces of u on the boundary, and, in the same time, the nullity of u.

Let W^* be the closure of $\mathcal{W}(\Omega_-)$ in $\mathbf{H}^1(\Omega_-)$; clearly, W^* is included in $W(\Omega_-)$. To prove that W^* is precisely $W(\Omega_-)$, let u be an element of $W(\Omega_-)$ that is orthogonal to W^*, or equivalently to $\mathcal{W}(\Omega_-)$; then, for any v in $\mathcal{W}(\Omega_-)$,

$$(u_1, v_1) + (u_2, v_2) + \left(\frac{\partial u_1}{\partial x_1}, \frac{\partial v_1}{\partial x_1}\right) + \left(\frac{\partial u_1}{\partial x_2}, \frac{\partial v_1}{\partial x_2}\right) + \left(\frac{\partial u_2}{\partial x_1}, \frac{\partial v_2}{\partial x_1}\right) + \left(\frac{\partial u_2}{\partial x_2}, \frac{\partial v_2}{\partial x_2}\right) = 0 \tag{2.7}$$

As v belongs to $\mathcal{W}(\Omega_-)$, there is a C^∞ function φ such that

$$v_1 = \frac{\partial \varphi}{\partial x_2}, \qquad v_2 = -\frac{\partial \varphi}{\partial x_1}, \tag{2.8}$$

v vanishes outside of a compact set of Ω_-. Conversely, any φ in $\mathcal{D}(\Omega_-)$ defines a v in $\mathcal{W}(\Omega_-)$, with the help of (2.4). Similarly, the theory of Beppo-Levi spaces shows that there exists ψ in $L^2_{\text{loc}}(\Omega_-) \subset \mathcal{S}'(\Omega_-)$ such that

$$u_1 = \frac{\partial \psi}{\partial x_2}, \qquad u_2 = -\frac{\partial \psi}{\partial x_1}. \tag{2.9}$$

If we substitute u and v in (2.7) with their expressions in terms of φ and ψ, we obtain:

$$\left(\frac{\partial \psi}{\partial x_2}, \frac{\partial \varphi}{\partial x_2}\right) + \left(\frac{\partial \psi}{\partial x_1}, \frac{\partial \varphi}{\partial x_1}\right) + 2\left(\frac{\partial^2 \psi}{\partial x_1 \partial x_2}, \frac{\partial^2 \varphi}{\partial x_1 \partial x_2}\right) + \left(\frac{\partial^2 \psi}{\partial x_2^2}, \frac{\partial^2 \varphi}{\partial x_2^2}\right) + \left(\frac{\partial^2 \psi}{\partial x_1^2}, \frac{\partial^2 \varphi}{\partial x_1^2}\right) = 0, \quad \forall \varphi \in \mathcal{D}(\Omega_-). \tag{2.10}$$

Therefore, in the sense of distributions, ψ satisfies the equation

$$-\Delta \psi + \Delta^2 \psi = 0 \text{ in } \Omega_-. \tag{2.11}$$

If we perform a Fourier transform in the tangential variable x_2, which is sent to k, it is not difficult to see that the general solution of (2.11) is of the form

$$\hat{\psi}(x_1, k) = \alpha \exp(|k|x_1) + \beta \exp((1+|k|^2)^{1/2} x_1) + \gamma \exp(-|k|x_1) + \delta \exp(-(1+|k|^2)^{1/2} x_1).$$

For $\hat{\psi}$ to be a temperate distribution, the coefficients γ and δ must vanish, because the corresponding modes increase exponentially at $x_1 = -\infty$. Therefore, $\hat{\psi}$ is of the form

$$\hat{\psi}(x_1, k) = \alpha \exp(|k|x_1) + \beta \exp(x_1 \sqrt{1+|k|^2}). \tag{2.12}$$

We will prove that α and β vanish identically. For this purpose, we need some regularity. We integrate $|k\hat{\psi}(x_1, k)|^2 = |\hat{u}_1|^2$ on Ω_-, and we obtain an estimate on α and β.

$$\int |k|^2 \left\{ \frac{|\alpha(k)|^2}{|k|} + \frac{|\beta(k)|^2}{\sqrt{1+|k|^2}} + 4 \operatorname{Re} \frac{\alpha(k)\overline{\beta(k)}}{|k|+\sqrt{1+|k|^2}} \right\} dk < +\infty. \tag{2.13}$$

We compute the smallest eigenvalue of the quadratic form that appears in (2.13). This shows that there is a positive weight $h(k)$ such that

$$\int |\hat{u}_1|^2 \, dk \, dx_1 \geqq \int |k|^2 h(k)[|\alpha|^2 + |\beta|^2](k) \, dk,$$

the weight h satisfies the estimates:

$$h(k) \simeq \tfrac{1}{2} \quad \text{in a neighborhood of } k = 0$$

$$h(k) \simeq \frac{1}{16|k|^5} \quad \text{in a neighborhood of } |k| = +\infty.$$

Therefore,

$$ik\alpha \text{ and } ik\beta \text{ belong to the space } H^{-5/2}. \tag{2.14}$$

If we return to formulation (2.10), written in Fourier variables, we obtain the variational problem

$$-(|k|^2\hat{\psi}, \bar{\hat{\varphi}}) + \left(\frac{\partial\hat{\psi}}{\partial x_1}, \frac{\partial\bar{\hat{\varphi}}}{\partial x_1}\right) + (|k|^4\hat{\psi}, \bar{\hat{\varphi}}) + \left(|k|^2\frac{\partial\hat{\psi}}{\partial x_1}, \frac{\partial\bar{\hat{\varphi}}}{\partial x_1}\right) + \left(\frac{\partial^2\hat{\psi}}{\partial x_1^2}, \frac{\partial^2\bar{\hat{\varphi}}}{\partial x_1^2}\right) = 0$$
$$\forall \varphi \in \mathcal{D}(\Omega_-). \tag{2.15}$$

According to (2.14) and (2.12), $\partial\hat{\psi}/\partial x_1$, $\partial^2\hat{\psi}/\partial x_1^2$ and $\partial^3\hat{\psi}/\partial x_1^3$ have a trace on Γ, even if it is in a very weak sense. Accordingly, we can perform several integrations by parts on (2.15), and we obtain the relations

$$\left[\frac{\partial\hat{\psi}}{\partial x_1} + |k|^2\frac{\partial\hat{\psi}}{\partial x_1} - \frac{\partial^3\hat{\psi}}{\partial x_1^3}\right]\Bigg|_\Gamma = 0$$
$$\left[\frac{\partial^2\hat{\psi}}{\partial x_1^2}\right]\Bigg|_\Gamma = 0,$$

substituting (2.12), we obtain a linear system:

$$|k|(1+|k|^2)\alpha + |k|^2\sqrt{1+|k|^2}\beta = 0$$
$$|k|^2\alpha + (1+|k|^2)\beta = 0.$$

This system is not singular, thus α and β vanish identically, and so does the potential ψ thanks to (2.12). This proves that u is zero, and that the orthogonal of the closure of $\mathcal{W}(\Omega_-)$ in $W(\Omega_-)$ is zero. □

We have an analogous result for the space $H(\Omega_-)$.

PROPOSITION 2.3. *The closure of $\mathcal{W}(\Omega_-)$ in $\mathbf{L}^2(\Omega_-)$ is precisely $H(\Omega_-)$.*

Proof. The technique of proof is absolutely identical to the technique employed for the previous proposition. It is in fact easier; the main steps begins with an analogue of equation (2.11):

$$\Delta\psi = 0 \quad \text{in } \Omega_-.$$

Then the Fourier transform of the potential y is of the form

$$\hat{\psi}(x_1, k) = \alpha \exp(|k|x_1),$$

and it is very easy to prove that $\alpha\sqrt{\gamma|k|}$ is square integrable; the remainder of the proof is left to the reader. □

Let us consider now the trace spaces of $W(\Omega_-)$ and $H(\Omega_-)$. The main result we need is as follows.

PROPOSITION 2.4. *For any $u=(u_1, u_2)^T$ in $W(\Omega_-)$, the trace of u_1 on Γ is in $H^{1/2}(\Gamma)$ and satisfies*

$$\int_{\mathbb{R}} |\hat{u}_1(0, k)|^2 \sqrt{|k|^2+\frac{1}{|k|^2}}\, dk < +\infty. \tag{2.16}$$

If $W^{1/2}$ is the space of functions that satisfy (2.16), *then the trace mapping $u \to u_1|_\Gamma$ from $W(\Omega_-)$ to $W^{1/2}$ is onto.*

Proof. Let u belong to $\mathcal{W}(\Omega_-)$; we may write

$$\frac{\partial}{\partial x_1}\left(\frac{|\hat{u}_1(x_1, k)|^2}{|k|}\right) = \frac{2}{|k|}\operatorname{Re}\left(u_1(x_1, k)\,\overline{\frac{\partial \hat{u}_1(x_1, k)}{\partial x_1}}\right)$$

$$= \frac{2}{|k|}\operatorname{Re}\left(u_1(x_1, k)\, ik\,\overline{\hat{u}_2(x_1, k)}\right),$$

since u is divergence free. The function u vanishes for x_1 small enough, because the support of u is compact. Thus we obtain

$$\int \frac{|\hat{u}_1(x_1, k)|^2}{|k|}\, dk \leqq \|u\|^2. \tag{2.17}$$

The trace space of $W(\Omega_-)$ is included in $\mathbf{H}^{1/2}(\Gamma)$; together with (2.17), we obtain the first statement of the proposition, because $\mathcal{W}(\Omega_-)$ is dense in $W(\Omega_-)$. To see that the trace mapping is onto, take the orthogonal of the image of $W(\Omega_-)$ in $W^{1/2}$ by the normal trace mapping. For all φ in $\mathcal{D}(\Omega_-)$, we have

$$\operatorname{Re}\left(\int_{\mathbb{R}} \left(|k|^2+\frac{1}{|k|^2}\right)^{1/2} \hat{u}_1 ik\, \bar{\hat{\varphi}}\, dk\right) = 0,$$

and this shows immediately that the mapping is onto. □

From the proof of Proposition 2.4, we deduce

COROLLARY 2.5. *For any u in $H(\Omega_-)$, the trace of u on Γ exists and satisfies*

$$\int_{\mathbb{R}} \frac{|\hat{u}_1(x_1, k)|^2}{|k|}\, dk < +\infty.$$

Moreover, the expression

$$s(u, v) = (u, v) + \int_{\mathbb{R}} \frac{\hat{u}_1(x_1, k)\,\overline{\hat{v}_1(x_1, k)}}{|k|}\, dk \tag{2.18}$$

is a scalar product on $H(\Omega_-)$, which is equivalent to the scalar product $(\cdot, \cdot)$ induced by $L^2(\Omega_-)$.

Proof. It is clear that

$$s(u, u) \geqq (\|u\|_0)^2;$$

the inequality

$$s(u, u) \leqq 2(\|u\|_0)^2$$

follows from (2.17). □

The dual space of $W^{1/2}$ will be denoted $W^{-1/2}$ and is the space of functions which satisfy

$$u \in W^{-1/2} \Leftrightarrow \int_{\mathbb{R}} |\hat{u}_1(0,k)|^2 \left(|k|^2 + \frac{1}{|k|^2}\right)^{1/2} dk < +\infty. \tag{2.19}$$

Remark 2.6. The space $W^{1/2}$ is included in $H^{1/2}$ because the weight $(|k|^2 + 1/|k|^2)^{1/2}$ satisfies the inequality

$$\left(|k|^2 + \frac{1}{|k|^2}\right)^{1/2} \geqq \frac{1+|k|^2}{2} \quad \forall k \in \mathbb{R}^*.$$

Dually, $W^{-1/2}$ contains $H^{-1/2}$.
This completes our review of divergence-free functional spaces.

2.3. Oseen system in full space. Consider the Navier-Stokes system in $\mathbb{R}^2 \times \mathbb{R}^+$:

$$u_t + (u \cdot \nabla)u - \nu \Delta u + \nabla p = 0; \nabla \cdot u = 0. \tag{2.20}$$

Here ∇ is the gradient operator, Δ is the Laplacian operator; $v \cdot \nabla$ denotes the differential operator $v_1 \partial/\partial x_1 + v_2 \partial/\partial x_2$.

We linearize this system around a constant state $a = (a_1, a_2)$, with a_1 positive, and we obtain a Stokes system with an advection term or Oseen system:

$$u_t + (a \cdot \nabla)u - \nu \Delta u + \nabla p = 0; \tag{2.21}$$

$$\nabla \cdot u = 0. \tag{2.22}$$

A differential operator $\mathcal{A}$ is defined by

$$\mathcal{A}(u, p) = u_t + (a \cdot \nabla)u - \nu \Delta u + \nabla p. \tag{2.23}$$

For functional analysis reasons, it will be convenient to study the differential operator $\mathcal{A}_\mu$ defined, for $\mu > 0$, by

$$\mathcal{A}_\mu(u, p) = u_t + \mu u + (a \cdot \nabla)u - \nu \Delta u + \nabla p. \tag{2.23}_\mu$$

We define a bilinear form $\mathbf{a}$ on $\mathbf{H}^1(\Omega)$ by

$$a(u, v) = ((\mathbf{a} \cdot \nabla)u, v) + \nu(\nabla u, \nabla v), \tag{2.24}$$

where

$$(\nabla u, \nabla v) = (\nabla u_1, \nabla v_1) + (\nabla u_2, \nabla v_2).$$

Clearly, $\mathbf{a}$ is continuous on $\mathbf{H}^1(\Omega)$; moreover, we have

$$((a \cdot \nabla)u, u) = \frac{1}{2} \int_\Omega a \cdot \nabla(|u|^2)\, dx = \frac{1}{2} \int_\Gamma a \cdot n |u|^2\, d\Gamma.$$

If $\Omega = \mathbb{R}^2$, this expression vanishes; if $\Omega = \Omega_-$, this expression is greater than or equal to zero. Therefore,

$$\mathbf{a}(u, u) \geqq \nu(\|\nabla u\|_0)^2;$$

there exists a positive constant α such that

$$(\|u\|_0)^2 + \mathbf{a}(u, u) \geqq \alpha(\|u\|_1)^2, \quad \forall u \text{ in } \mathbf{H}^1(\Omega). \tag{2.25}$$

We will need a partially antisymmetrized form of $\mathbf{a}$, in order to uncouple the boundary part of Green's formula; let

$$\tilde{\mathbf{a}}(u, v) = \tfrac{1}{2}[((a \cdot \nabla)u, v) - ((a \cdot \nabla)v, u)] + \nu(\nabla u, \nabla v). \tag{2.26}$$

As a is constant, $\mathbf{a}$ and $\tilde{\mathbf{a}}$ can differ only by a boundary term; if $\Omega = \mathbb{R}^2$, there is no such term; if $\Omega = \Omega_-$,

$$\tilde{\mathbf{a}}(u, v) - \mathbf{a}(u, v) = -\frac{1}{2}\int_\Gamma a_1 u \cdot v \, d\Gamma. \tag{2.27}$$

In particular, in the full space case,

$$\mathbf{a}(u, u) = \tilde{\mathbf{a}}(u, u) = \nu \int |\nabla u|^2 \, dx. \tag{2.28}$$

Consider now the linearized Navier-Stokes system in the plane:

$$\mathcal{A}_\mu(u, p) = f \quad \text{in } \mathbb{R}^2 \times \mathbb{R}^+; \tag{2.29}$$

$$\nabla \cdot u = 0 \quad \text{in } \mathbb{R}^2 \times \mathbb{R}^+; \tag{2.30}$$

$$u(\cdot, 0) = u^0 \quad \text{in } \mathbb{R}^2. \tag{2.31}$$

This system is well posed, if we take u^0 and f in adequate functional spaces.

PROPOSITION 2.7. *For all u^0 in $H(\mathbb{R}^2)$, and all f in $L^2(0, \infty; L^2(\mathbb{R}^2))$, and for all nonnegative μ, there exists a unique u and a p unique up to an additive constant such that* (2.29)-(2.31) *hold and*

$$u \in L^\infty_{\text{loc}}([0, \infty); H(\mathbb{R}^2)),$$
$$\nabla u \in L^2_{\text{loc}}([0, \infty); \mathbf{L}^2(\mathbb{R}^2)),$$
$$u_t \in L^2_{\text{loc}}([0, \infty); V'(\mathbb{R}^2)),$$
$$p = \frac{\partial P}{\partial t},\ P \in L^2_{\text{loc}}([0, \infty); BL(H^{-1}(\mathbb{R}^2))).$$

Here $V'(\mathbb{R}^2)$ is the dual of $V(\mathbb{R}^2)$ and $BL(H^1(\mathbb{R}^2))$ is the Beppo-Levi space of functions whose gradient is in $H^1(\mathbb{R})$. The function u belongs to $L^p_{\text{loc}}([0, \infty); X)$ if any restriction of u to a finite time interval $[0, T]$ belongs to $L^p([0, T]; X)$.

Moreover, for any strictly positive μ, the following global estimates hold:

$$u \in L^\infty([0, \infty); H(\mathbb{R}^2)),$$
$$\nabla u \in L^2([0, \infty); L^2(\mathbb{R}^2)),$$
$$u_t \in L^2([0, \infty); V'(\mathbb{R}^2)),$$
$$p = \frac{\partial P}{\partial t},\ P \in L^2([0, \infty); BL(H^{-1}(\mathbb{R}^2))).$$

Remark 2.8. If Ω were a bounded open set with smooth enough boundary, it would be an exercise to extend the existence proof in [28] for the Stokes problem to the present situation. Though the forthcoming proof is quite classical, it does not appear to be written with all the necessary details in the literature of which we know.

Proof. We know from [24] that for every positive and finite T, there exists a unique u in $L^2(0, T; V(\mathbb{R}^2))$ with $\partial u/\partial t$ in $L^2(0, T; V'(\mathbb{R}^2))$ such that

$$(u_t, v) + \mathbf{a}(u, v) + \mu(u, v) = (f, v) \quad \forall v \in V(\mathbb{R}^2); \tag{2.32}_\mu$$

$$u(0) = u^0. \tag{2.33}$$

Let us first consider the case $\mu = 0$. Relation (2.28) implies an energy estimate

$$\frac{1}{2}\frac{d}{dt}\|u\|_0^2 + \nu\|\nabla u\|_0^2 \leqq \|f(t)\|_0 \|u\|_0.$$

From this inequality we make a classical Gronwall estimate and deduce the estimates on u and ∇u, using the coerciveness; the variational formulation $(2.32)_\mu$ gives the estimate on u_t. Precisely, we get the following estimates

$$\|u(t)\|_0 \leqq \|u^0\|_0 + \sqrt{t}\|f\|_{L^2([0,t]\times\mathbb{R}^2)},$$

$$\int_0^t (\|\nabla u\|_0)^2\, ds \leqq C(\|f\|_{L^2([0,t]\times\mathbb{R}^2)})(\|u^0\|_0 + \|f\|_{L^2([0,t]\times\mathbb{R}^2)})(1+t).$$

These two estimates show polynomial growth in time. In order to have an estimate on p, we proceed as in [28]. Let

$$U(t) = \int_0^t u(s)\, ds, \qquad F(t) = \int_0^t f(s)\, ds.$$

Then ∇U belongs to $L^2(0, T; L^2(\mathbb{R}^2))$. If we integrate (2.21) in time, we obtain

$$(u(t) - u^0 + (a\cdot\nabla)U - \nu\Delta U, v) = (F, v) \quad \forall v \text{ in } V(\mathbb{R}^2).$$

The expression $u(t) - u^0 + (a\cdot\nabla)U - \nu\Delta U - F$ is orthogonal to divergence-free vectors, and belongs to $L^\infty(0, T; \mathbf{H}(\mathbb{R}^2)) + L^2(0, T; \mathbf{H}^{-1}(\mathbb{R}^2))$, which is included in $L^2(0, T; \mathbf{H}^{-1}(\mathbb{R}^2))$. Therefore, there exists P in $L^2(0, T; BL(H^{-1}(\mathbb{R}^2))$ such that

$$u(t) - u^0 + (a\cdot\nabla)U - \nu\Delta U + \nabla P = F.$$

If we define $p = \partial P/\partial t$, we obtain the announced estimate.

For the last statement, we observe that (u, p) is a solution of $(2.32)_0$ if and only if $(w, q) = (ue^{-\mu t}, pe^{-\mu t})$ is solution of $(2.32)_\mu$.

Thus, the polynomial growth estimates for u ensure the global estimates for w. □

There is a regularity result which will be useful in what follows; denote

$$H^\infty(\mathbb{R}^2) = \bigcap_{m\geqq 0} H^m(\mathbb{R}^2); \tag{2.34}$$

this is a Frechet space, with an obvious topology. The regularity result is as follows:

LEMMA 2.9. *Let u^0 belong to $H^\infty(\mathbb{R}^2)$, and f to $C^\infty(\mathbb{R}^+; H^\infty(\mathbb{R}^2))$. Then, the solution (u, p) of (2.29)-(2.31) belongs to $C^\infty(\mathbb{R}^+; H^\infty(\mathbb{R}^2))$.*

Proof. The spatial derivatives of u are divergence free, and satisfy (2.24); if we apply the estimates of Proposition 2.7 to the differentiated equation, we obtain the desired estimates. In order to differentiate in time, we check that the time derivative $u_t(., t)$ is in $H(\mathbb{R}^2)$; from (2.29), $u_t(., t)$ is the projection of $\nu\Delta u - (a\cdot\nabla)u$ onto $H(\mathbb{R}^2)$; u_t satisfies (2.29), (2.30), and thus Proposition 2.6 is applicable. By induction,

$$u \in H^m(0, T; H^m(\mathbb{R}^2)), \quad \forall\, m \text{ in } \mathbb{N}.$$

Leaving all details to the reader, this ends the proof. □

Remark 2.10. For μ strictly positive, it is possible to prove that if u^0 belongs to $H^\infty(\mathbb{R}^2)$, and f belongs to $\mathcal{S}([0,\infty); H^\infty(\mathbb{R}^2))$, then u belongs to $\mathcal{S}([0,\infty); H^\infty(\mathbb{R}^2))$. This result will be proved indeed for the case of the half-plane in the next section. The reader is referred to Proposition 3.1, whose proof can be completely copied to obtain this result. In any case, it is a question of estimating solutions of linear equations with a nice exponentially decreasing behavior, and the proof is an easy application of semigroup theory.

Remark 2.11. The solution of (2.29)-(2.31) does not decrease fast at infinity in space, in general; the trouble is with the pressure. Assume that f vanishes for large x; by taking the divergence of (2.29), in the smooth case, one can see that $\Delta p = \nabla\cdot f$, so that the pressure is harmonic; if p decreased rapidly at infinity in space to a constant,

it would be identically equal to this constant (see Appendix B for a proof of this result); this is not a general situation. If ∇p decays like some power of x, at infinity, then, once ∇p is known, u is essentially a solution of the heat equation with source $\nabla p + f$, and it cannot, in general, decrease rapidly at infinity in space.

2.4. On specific trace results for the solution of Oseen system. Consider the solution u of (2.29)-(2.31). Extend u and p by 0 for $t \leqq 0$, and denote the extended functions still by u and p. Then,

$$\mathcal{A}_\mu(u, p) = u^0 \otimes \delta^t + f \text{ in } \mathbb{R}^2 \times \mathbb{R};\ \nabla \cdot u = 0. \tag{2.35}$$

We cannot expect that u restricted to $\sum = \{(0, x_1, t)/(x_1, t) \in \mathbb{R}^2\}$ will be smooth in time, even if we assume that

(2.36) The support of u^0 is compact and included in Ω_-,

(2.37) The support of f is in the product of a compact subset of Ω_- with $\mathbb{R}_+$.

Remark 2.12. We explain why smoothness in time cannot be expected in general, and give sufficient conditions to have it. Denote by Π the projection in $\mathbf{L}^2(\mathbb{R}^2)$ onto $H(\mathbb{R}^2)$; assume that u^0 is smooth enough for the foregoing computations. Then, interpreting the pressure as a Lagrange multiplier, we can take a limit as t decreases to zero:

$$u_t(\cdot, 0^+) = \Pi f + u^0 \otimes \delta^t - (a \cdot \nabla) u^0 + \nu \Delta u^0 - \mu u^0.$$

In this relation all the terms containing u^0 vanish on $\sum \cap \{t = 0\}$, thanks to assumption (2.36). Nevertheless, there is no reason why Πf should vanish there, since Π is not a local operator. Thus, in order to have some smoothness in time, we should ask that f, and a number of its time derivatives vanish at time 0. This assumption is not reasonable, and we shall not make it because it turns out that we are interested in less than the regularity of u.

More precisely, let $\sigma = \operatorname{sgn}(k)$ and let $\mathcal{H}$ denote the Hilbert transform on Γ [27, Chap. V and VI]:

$$(\mathcal{H}u)\hat{}(k) = -i\sigma \hat{u}(k). \tag{2.38}$$

In physical variables, the Hilbert transform is defined as the convolution with the principal value $v.p.\,(1/\pi x_2)$. We shall see later that we are interested only in the trace of $u_1 + \mathcal{H}u_2$ on Σ. It turns out that this trace is very regular in t and x_2, under the support conditions (2.36), (2.37).

A sequence of lemmas will describe the precise regularity of the trace of $u_1 + \mathcal{H}u_2$.

LEMMA 2.13. *Let u_0 belong to $\mathbf{H}(\mathbb{R}^2)$, and f to $L^2(0, \infty; \mathbf{H}(\mathbb{R}^2))$, and let u be the solution of* (2.29)-(2.31). *Assume* (2.36) *and* (2.37). *Let*

$$z = u_1 + \mathcal{H}u_2. \tag{2.39}$$

Then z belongs to $L^2_{\text{loc}}([0, \infty); H^1(\mathbb{R}^2)) \cap L^\infty_{\text{loc}}([0, \infty); L^2(\mathbb{R}^2))$ and satisfies a heat equation of the form

$$z_t + \mu u + a \cdot \nabla u - \nu \Delta u = g, \tag{2.40}$$

with an initial condition

$$z(\cdot, 0) = u_1(\cdot, 0) + \mathcal{H}u_2(\cdot, 0) \tag{2.41}$$

belonging to $L^2(\mathbb{R}^2)$. Here, g has support in $(-\infty, -X) \times \mathbb{R} \times \mathbb{R}^+$, and X is some strictly positive number.

Proof. The function z is well defined as a function of x_1, x_2, and t, because $\mathcal{H}$ is an isometry from $L^2(\mathbb{R})$ to itself. Moreover, z belongs to the spaces mentioned in the lemma, because $\mathcal{H}$ commutes with the differentiations, so that if w is in $H^1(\mathbb{R}^2)$, so is $\mathcal{H}w$. If we compute the right-hand side of (2.40) in the sense of distributions, we obtain

$$g = f_1 + \mathcal{H}f_2 - \frac{\partial p}{\partial x_1} - \mathcal{H}\frac{\partial p}{\partial x_2}.$$

It remains to show that $\partial p/\partial x_1 + \mathcal{H}(\partial p/\partial x_2)$ has its support in $\{x_1 \leqq -X\}$, because assumption (2.37) shows that f has its support in this set. If we take the divergence of (2.29) in the sense of distributions, we have

$$\Delta p = \nabla \cdot f.$$

Therefore, p is harmonic in the region $\{x_1 \geqq -X\} \times \mathbb{R} \times \mathbb{R}^+$. By partial Fourier transform in x_2,

$$\frac{\partial^2 \hat{p}}{\partial x_1^2} - |k|^2 \hat{p} = 0.$$

As p is temperate in x_1, x_2, p is necessarily of the form, for $x_1 \geqq -X$,

$$\hat{p} = \hat{p}(0, k, .)\, e^{-|k|x_1},$$

and therefore,

$$\frac{\partial \hat{p}}{\partial x_1} + \left(\mathcal{H}\frac{\partial p}{\partial x_2}\right)^{\wedge} = (-|k| + ik(-i\sigma))\hat{p} = 0 \text{ for } x_1 \geqq -X.$$

This proves that the support of g is indeed in the region $\{x_1 \leqq -X\}$. □

LEMMA 2.14. *Let w be a solution of*

$$w_t + a_2 \partial w/\partial x_2 - \nu \Delta w = g \quad \text{for } x \in \mathbb{R}^2, \quad t \in \mathbb{R},$$

$$w = 0 \quad \text{for } t \leqq T_1,$$

where the support of the distribution g is included in $(-\infty, -X] \times \mathbb{R} \times \mathbb{R}^+$. If g belongs to $H^s(\mathbb{R}^2 \times \mathbb{R})$, then the trace of w on $\Sigma = \{0\} \times \mathbb{R} \times \mathbb{R}$ belongs to $H^\infty(\mathbb{R} \times [0, T])$, for all T.

Proof. Among the many possible ways of proving this result, we choose the one which is the closest to the spirit of this article; namely, we consider the problem with respect to the variable x_1, instead of a problem in time. We will perform a Fourier transform in time and in the partial space variable x_2 and perform a Fourier analysis of the ordinary differential equation obtained in this fashion. For a correct argument, we multiply w by $e^{-\mu t}$, where μ is a strictly positive number. Then $v = w\, e^{-\mu t}$ satisfies the equation

$$v_t + \mu v + a_2 \frac{\partial v}{\partial x_2} - \Delta v = g\, e^{-\mu t} = h, \tag{2.42}$$

where h has the same support properties and smoothness properties as g. This amounts to performing a Fourier–Laplace transform in time instead of a Fourier transform. If ω denotes the dual variable of t, and k the dual variable of x_2, the partial Fourier transform of (2.42) is

$$(i\omega + \mu + ia_2 k + \nu|k|^2)\hat{v} - \nu \frac{\partial^2 \hat{v}}{\partial x_1^2} = \hat{h}. \tag{2.43}$$

Let ρ be the root of $\nu\rho^2-(i(\omega+a_2k)+\mu+\nu|k|^2)=0$ which has positive real part. An elementary computation shows that the unique solution of (2.43) which is temperate is given by

$$\hat{v}(x_1, k, \omega) = -\frac{1}{2\rho\nu}\left[2\int_{-\infty}^{x_1} \operatorname{sh}(\rho(y-x_1))\hat{h}(y, k, \omega)\, dy\right] + \int_{-\infty}^{\infty} \exp(\rho(x_1-y))\hat{h}(y, k, \omega)\, dy.$$

Thus, we can write

$$\hat{v}(0, k, \omega) = -\frac{1}{2\rho\nu}\left[2\int_{-\infty}^{-X} \operatorname{sh}(\rho y)\hat{h}(y, k, \omega)\, dy + \int_{-\infty}^{-X} \exp(-\rho y)\hat{h}(y, k, \omega)\, dy\right]$$
$$= \frac{1}{2\rho\nu}\int_{-\infty}^{X} e^{\rho y}\hat{h}(y, k, \omega)\, dy.$$

We can see that

$$\operatorname{Re}(\rho) \geqq C(1+|k|+\sqrt{|\omega|}), \tag{2.44}$$

for some constant C strictly positive, and thus, it is possible to estimate

$$\int (1+|k|^2+|\omega|^2)^m |\hat{v}(0, k, \omega)^2\, dk\, d\omega$$
$$\leqq \int (1+|k|^2+|\omega|^2)^m \int_{-\infty}^{-X} |e^{\rho y}\hat{h}(x_1, k, \omega)\, dy|^2\, dk\, d\omega$$
$$\leqq \int (1+|k|^2+|\omega|^2)^m \left\{\int_{-\infty}^{X} |e^{2\rho y}|\, dy\right\}\left\{\int |\hat{h}(x, k, \omega)|^2\, dx\right\} dk\, d\omega.$$

If we first strengthen somewhat the hypotheses by assuming that, for some n,

$$\int (1+|k|^2+|\omega|^2)^n |\hat{h}(x, k, \omega)|^2\, dx\, dk\, d\omega < \infty,$$

then, the result we look for can be easily deduced:

$$\int_{-\infty}^{-X} |e^{2\rho y}|\, dy = \frac{e^{-2X \operatorname{Re}(\rho)}}{2 \operatorname{Re}(\rho)},$$

and thanks to (2.24), this quantity is dominated by all powers of k and ω, and the result is proved. To get rid of the extra assumption we made, we observe that an h in $H^s(\mathbb{R}^2\times\mathbb{R})$ is a finite sum of x_1 derivatives of some functions h_j, and the h_j can be taken with the same support property as h, and each h_j satisfies

$$\int_{-\infty}^{-X} (1+|k|^2+|\omega|^2)^n |\hat{h}_j(x, k, \omega)|^2\, dx\, dk\, d\omega < \infty.$$

We treat the terms

$$\int_{-\infty}^{-X} e^{\rho y}\frac{\partial^j \hat{h}_j(y, k, \omega)}{\partial y^j}\, dy$$

by performing an integration by parts, which will amount to a multiplication by some extra powers of ρ; these powers will be nevertheless dominated by $\exp(2X \operatorname{Re}(\rho))$, and

$$\int (1+|k|^2+|\omega|^2)^m \left| \int_{-\infty}^{-X} e^{\rho y} \frac{\partial^j \hat{h}_j(x_1, k, \omega)}{\partial y^j} dy \right|^2 dk\, d\omega$$

is finite for all real m.

Going back to w, we obtain the required result. □

To conclude this sequence of results, we can state now the

LEMMA 2.15. *Let z be as in Lemma 2.13. Then, the trace of z on Σ belongs to $H^\infty(\Sigma)$.*

Proof. Consider the new variables

$$t' = t,\ x_1' = x_1 - a_1 t,\ x_2' = x_2.$$

In these new variables, the equation satisfied by z becomes

$$z_{t'} + a_2 \frac{\partial z}{\partial x_2} - \nu \Delta' z = g,$$

and the support of g is included in the set

$$\{(x', t')/t' \geqq 0 \text{ and } x_1' + a_1 t' \leqq -X\}.$$

As a_1 is strictly positive, $a_1 t'$ is less than or equal to zero, and we are in the case of Lemma 2.14. □

3. Analysis of the transparent boundary condition.

3.1. Introduction. Let (u, p) be the solution of (2.29)-(2.31) with initial data satisfying (2.36)-(2.37). The normal constraint σ_n is defined by

$$\sigma_n = (\sigma_{11}, \sigma_{12}); \quad \sigma_{11} = \nu \frac{\partial u_1}{\partial x_1} - p; \quad \sigma_{12} = \nu \frac{\partial u_2}{\partial x_1}. \tag{3.1}$$

We will show in this section that the restrictions to $\Sigma = \{0\} \times \mathbb{R} \times \mathbb{R}^+$ of u and the normal constraint σ_n satisfy a linear pseudo-differential relation. This relation will be computed by Fourier techniques, working on the problem

$$\begin{cases} a(u, p) = 0 & \text{in } \Omega_+ \times \mathbb{R}; \\ \nabla \cdot u = 0 & \text{in } \Omega_+ \times \mathbb{R}; \\ u(0, x_2, t) = g(x_2, t) & \text{in } \mathbb{R} \times \mathbb{R}. \end{cases} \tag{3.2}$$

We assume here that g is given in $\mathcal{S}(\mathbb{R}; \mathbf{H}^\infty(\mathbb{R}))$. The choice of $\mathbf{H}^\infty(\mathbb{R})$ is justified by Remark 2.11. On the other hand, there is no such constraint in time, and it is permissible to have a solution with values in $\mathbf{H}^m(\mathbb{R})$ which decreases fast in time for all nonnegative m. Moreover, we ask that

$$k \to \frac{1}{\sqrt{|k|}} \hat{g}(k, \omega) \in \mathbf{L}^2(\mathbb{R}), \quad \forall\ \omega.$$

Under these assumptions, we will show at Proposition 3.1 that for all such g, (3.2) admits a unique solution u in $C^\infty(\Omega_+ \times \mathbb{R})$.

The mapping

$$\mathcal{E} : g \to \sigma_n$$

can be completely described in Fourier variables:

$$(\mathcal{E}g)\hat{}(k, \omega) = E(k, \omega) \hat{g}(k, \omega), \tag{3.3}$$

where E is a two-by-two matrix that will be given explicitly in terms of k and ω at Corollary 3.3.

From this operator $\mathscr{E}$, we will define another operator $\mathscr{L}$, which is a nice pseudo-differential operator, such that the restriction of the solution of Oseen system to the left half-plane Ω_- satisfies the variational equation

$$s(u_t, v) + \tilde{\mathbf{a}}(u, v) + \langle \mathscr{L}g, v|\Sigma\rangle = 0, \tag{3.4}$$

where s is the scalar product defined at (2.18) and $\mathscr{L}$ is defined with the help of an explicit matrix L by

$$(\mathscr{L}g)\hat{}(k, \omega) = L(k, \omega)\hat{g}(k, \omega).$$

The next step is to study the properties of $\mathscr{L}$. This operator is causal, which means that if g vanishes for $t \leqq 0$, so does $\mathscr{L}g$. This property plus some estimates will enable us to extend $\mathscr{L}$ to much larger spaces.

In view of well-posedness results, we shall prove that $\mathscr{L}$ has some useful positivity properties; in particular, the symmetrized matrix $(L+L^*)/2$ is positive semidefinite.

With this detailed study of $\mathscr{L}$, we prove an existence and uniqueness result for the solution of

$$\mathscr{A}(u, p) = 0 \quad \text{in } \Omega_- \times \mathbb{R}, \tag{3.5}$$

$$\nabla \cdot u = 0 \quad \text{in } \Omega_- \times \mathbb{R}, \tag{3.6}$$

$$u(x_1, x_2, 0) = u^0(x_1, x_2) \quad \text{in } \Omega_-, \tag{3.7}$$

$$\sigma_n|_\Sigma = \mathscr{E}(u|_\Sigma). \tag{3.8}$$

This problem is written in variational form, and, with very smooth data, it admits a solution that is simply the restriction of the full-space problem with initial data extended by zero in Ω_+; the uniqueness will be a consequence of the positivity of the operators. Once we have the uniqueness, we can extend the class of solutions for which we have a solution, and, then, conditions (3.8) may be called transparent.

Most of the time, it will be convenient to replace (3.5) by $(3.5)_\mu$ where

$$\mathscr{A}_\mu(u, p) = \mathscr{A}(u, p) + \mu u; \tag{3.5}_\mu$$

here, μ is a positive number. This amounts to considering the system solved by $(u, p)e^{-\mu t}$, or, in other words, to extend the frequency ω to the half plane $Im(\omega) < 0$. This is permissible because we work with causal operators, and we can apply the Paley–Wiener–Schwarz theorem.

3.2. The boundary problem for Oseen system. In this section, we consider the problem

$$\begin{cases} \mathscr{A}_\mu(u, p) = 0 \quad \text{in } \Omega_+ \times \mathbb{R}, \\ \nabla \cdot u = 0 \quad \text{in } \Omega_+ \times \mathbb{R}, \\ u(0, x_2, t) = g(x_2, t) \quad \text{in } \mathbb{R} \times \mathbb{R}, \end{cases} \tag{3.9}$$

where

$$\mathscr{A}_\mu(u, p) = \partial u/\partial t + \mu u + (a \cdot \nabla)u - \nu \Delta u + \nabla p \tag{3.10}$$

and

$$\mu > 0.$$

If g vanishes for $t \leqq t_0$, and if (u, p) is a solution of (3.9) for $\mu > 0$, then $(ue^{\mu t}, pe^{\mu t})$ is a solution of (3.9) for $\mu = 0$, with data $g_\mu(x_2, t) = g(x_2, t)\, e^{\mu t}$.

We first prove a result of existence and regularity.

PROPOSITION 3.1. *Let Z be the subspace of $\mathscr{S}'(\mathbb{R})$ defined by*

$$g \in Z \text{ iff } g \in H^\infty(\mathbb{R}) \text{ and } \mathscr{F}^{-1}\left(\frac{\hat{g}_1(k)}{\sqrt{|k|}}\right) \in H^\infty(\mathbb{R}). \tag{3.11}$$

Assume that

$$g \in \mathscr{S}(\mathbb{R}; Z). \tag{3.12}$$

Then (3.9) *possesses a unique solution (u, p) such that,*

$$u \in L^\infty(\mathbb{R}; H(\Omega_+)), \tag{3.13}$$

$$\nabla u \in L^2(\mathbb{R}; \mathbf{L}^2(\Omega_+)), \tag{3.14}$$

$$\frac{\partial u}{\partial t} \in L^2(\mathbb{R}; W'(\Omega_+)), \tag{3.15}$$

$$\nabla p \in L^\infty(\mathbb{R}; \mathbf{L}^2(\Omega_+)) + L^2(\mathbb{R}; \mathbf{H}^{-1}(\Omega_+)). \tag{3.16}$$

Moreover, u is infinitely differentiable, and if $\mu > 0$,

$$u, p \in \mathscr{S}(\mathbb{R}; \mathbf{H}^\infty(\Omega_+)). \tag{3.17}$$

Proof. Let us first construct a function z such that

$$\begin{cases} z \in \mathscr{S}(\mathbb{R}; \mathbf{H}^\infty(\Omega_+)), \\ \nabla \cdot z = 0, \\ z|_\Sigma = g. \end{cases} \tag{3.18}$$

the function z will be a sum of two functions defined by different means. We first extend g_1; let ζ be defined by

$$\hat{\zeta}_1(x_1, k, \omega) = \hat{g}_1(k, \omega) \exp(-|k|x_1),$$
$$\hat{\zeta}_2(x_1, k, \omega) = -i\sigma \hat{g}_1(k, \omega) \exp(-|k|x_1),$$

where

$$\sigma = \operatorname{sign}(k).$$

We have used the Hilbert transform, of symbol $-i\sigma$, mentioned in § 2, and defined at (2.38):

$$\zeta_2(x_1, \cdot, t) = \mathscr{H}\zeta_1(x_1, \cdot, t).$$

Moreover,

$$\nabla \cdot \zeta = 0;\ \zeta_1(0, x_2, t) = g_1(x_2, t).$$

Now, in order to compensate for the bad boundary condition of the second component, we define a function h by

$$h(x_1, x_2, t) = \psi(x_1)[g_2(x_2, t) - (\mathscr{H}g_1)(x_2, t)],$$

where ψ belongs to $\mathscr{D}(\Omega_+)$, $\psi(0) = 0$, $\psi'(0) = 1$, and finally, we let

$$z_1 = \zeta_1 - \frac{\partial h}{\partial x_2}; \qquad z_2 = \zeta_2 + \frac{\partial h}{\partial x_1}.$$

Clearly h satisfies the boundary conditions. If (3.12) holds, then an integration in x_1 shows that

$$\int_0^\infty \int_{\mathbb{R}} \int_{\mathbb{R}} (1+|k|^2)^m |\hat{\zeta}_1(x_1, k, \omega)|^2 dx_1 dk\, d\omega$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} (1+|k|^2)^m (2|k|)^{-1} |\hat{g}_1(k, \omega)|^2\, dk\, d\omega < \infty.$$

Similarly,

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \int_0^\infty \left| \frac{\partial^m \hat{\zeta}_1(x_1, k, \omega)}{\partial x_1^m} \right|^2 dx_1\, dk\, d\omega = \int_{\mathbb{R}} \int_{\mathbb{R}} \int_0^\infty |k|^{2m} |\hat{\zeta}_1(x_1, k, \omega)|^2\, dx_1\, dk\, d\omega$$

$$= \frac{1}{2} \int_{\mathbb{R}} \int_{\mathbb{R}} |k|^{2(m-1/2)} |\hat{g}_1(k, \omega)|^2\, dk\, d\omega < \infty.$$

This shows that ζ_1 belongs to $L^2(\mathbb{R}; H^m(\Omega_+))$, for all m. We look at time derivatives multiplied by polynomials in t, and we show by induction that ζ_1 belongs to $\mathcal{S}(\mathbb{R}; H^\infty(\Omega_+))$. A similar argument shows that ζ_2 belongs to $\mathcal{S}(\mathbb{R}; H^\infty(\Omega_+))$. By construction, g_1 and g_2 belong to $\mathcal{S}(\mathbb{R}; H^\infty(\mathbb{R}))$; the Hilbert transform in the variable x_2 leaves this space invariant; therefore, $\mathcal{H}g_1$ belongs to it, and so h and its gradient belong to $\mathcal{S}(\mathbb{R}; H^\infty(\Omega_+))$. The function z we obtain finally satisfies (3.18).

Let now

$$u = v + z;$$

then (u, p) solves (3.9) if and only if (v, p) solves

$$\begin{cases} \mathcal{A}_\mu(v, p) = -\mathcal{A}_\mu(z, 0) & \text{in } \Omega_+ \times \mathbb{R}, \\ \nabla \cdot v = 0 & \text{in } \Omega_+ \times \mathbb{R}, \\ v(0, x_2, t) = 0 & \text{in } \mathbb{R} \times \mathbb{R}. \end{cases} \tag{3.19}$$

The existence of a solution of (3.19) is obtained by many classical methods; for instance, define the generator A_0 of the Stokes semigroup (without advection) by

$$D(A_0) = \{u \in H_0(\Omega_+) / v \to \int \nabla u \cdot \nabla v\, dx \text{ is continuous}\}$$

and

$$(A_0 u, v)_{H_0} = \nu \int \nabla u \cdot \nabla v\, dx. \tag{3.20}$$

The operator A_0 is defined starting from a quadratic form on the Hilbert space H_0; therefore, according to Phillips' theorem, it generates a contraction semigroup in H_0. Its domain can be computed explicitly and is equal to

$$D(A_0) = \left\{ u \in H_0(\Omega_+) \cap \mathbf{H}^2(\Omega_+) \Big/ \frac{\partial u_2}{\partial x_1}(0, x_2) = 0 \right\}.$$

Let A be defined by

$$D(A) = D(A_0);$$

$$(Au, v)_{H_0} = \int [(a \cdot \nabla) u \cdot v + \nu \nabla u \cdot \nabla v]\, dx. \tag{3.21}$$

Then A is obtained from A_0 by adding a strongly relatively bounded perturbation to A_0; this is because we have the inequality for all ε

$$\|\nabla u\|_0 \leqq \varepsilon \|\Delta u\|_0 + C(\varepsilon)\|u\|_0.$$

Therefore, A generates a strongly continuous semigroup $\mathcal{T}(t)$; this semigroup turns out to satisfy the estimate,

$$\|\mathcal{T}(t)\| \leqq e^{-\mu t}, \text{ for all } t \geqq 0,$$

because, for all u in $D(A)$, $(Au, u) \geqq \mu(\|u\|_0)^2$. The solution of (3.20) is given by

$$v(\cdot, t) = \int_{-\infty}^{t} \mathcal{T}(t-s)(-\mathcal{A}_\mu(z, 0)(s))\, ds,$$

and it is clear that $v(\cdot, t)$ is bounded uniformly on $\mathbb{R}$, with values in H_0, and that, if μ is positive, it decreases fast to zero as t tends to $\pm\infty$. All the other estimates are easy to obtain by differentiation and semigroups estimates, and they are global on $\mathbb{R}$. Details are left to the reader. □

We now perform the Fourier analysis of (3.9). For simplicity of notation, the frequency variable ω can be real, or complex with negative imaginary part; if needed, it will be denoted $\tau = \omega - i\mu$ when this negative imaginary part is present. We need a few notations; the differential operator C is given by

$$Cf = -\nu \frac{\partial^2 f}{\partial x_1^2} + a_1 \frac{\partial f}{\partial x_1} + (\nu|k|^2 + i\tau + a_2 ik) f; \tag{3.22}$$

its associated characteristic polynomial is given by

$$P(k, \tau; \lambda) = -\nu\lambda^2 + a_1\lambda + \nu|k|^2 + i\tau + a_2 ik. \tag{3.23}$$

The discriminant of P is given by:

$$a_1^2 + 4\nu(i(\tau + a_2 k) + \nu|k|^2). \tag{3.24}$$

The real part of (3.24) is positive for all ω and k, and for all $\mu \geqq 0$, because we assumed $a_1 > 0$. We denote by ρ the determination of the square root of (3.24) with positive real part:

$$\rho^2 = a_1^2 + 4\nu(i(\tau + a_2 k) + \nu|k|^2), \text{ Re } \rho > 0. \tag{3.25}$$

Then, the roots of $P(k, \tau, \cdot)$ are given by

$$\lambda = \frac{a_1 - \rho}{2\nu}; \quad \lambda' = \frac{a_1 + \rho}{2\nu},$$

and it is not difficult to check that

$$\text{Re } \lambda < 0 \text{ except if } k = \tau = 0. \tag{3.26}$$

Moreover, we have immediately the following important estimate; there exists a strictly positive constant γ such that

$$\frac{1}{\gamma}(1 + |k| + \sqrt{|\tau|}) \leqq |\rho| \leqq \gamma(1 + |k| + \sqrt{|\tau|}). \tag{3.27}$$

The first result pertaining to the Fourier analysis of (3.9) is as follows.

PROPOSITION 3.2. *Let (u, p) be the solution of* (3.9). *Then, there exist locally integrable functions α and β such that*

$$\hat{u}(x_1, k, \tau) = \exp(-|k|x_1)\alpha(k, \tau) + \exp(\lambda x_1)\beta(\alpha, \tau), \tag{3.28}$$

where

$$\text{(3.29)}\qquad \begin{cases} \alpha_1 = \dfrac{\lambda \hat{g}_1 + ik\hat{g}_2}{\lambda + |k|}, \beta_1 = k\dfrac{\sigma \hat{g}_1 - i\hat{g}_2}{\lambda + |k|}, \\ \alpha_2 = -i\sigma\dfrac{\lambda \hat{g}_1 + ik\hat{g}_2}{\lambda + |k|}, \beta_2 = i\lambda\dfrac{\sigma \hat{g}_1 - i\hat{g}_2}{\lambda + |k|}. \end{cases}$$

Moreover, any couple of distributions (α', β') *which satisfies* (3.27) *is equal to* (α, β), *up to the addition of* $(1, -1)S$, *where* S *is an arbitrary distribution with values in* $\mathbb{R}^2$ *and support in* $\{0\}_{k,\omega}$.

Proof. We perform a partial Fourier transform on u and p; the transformed quantities are denoted $\hat{u} = (\hat{u}_1, \hat{u}_2)$ and $\hat{p}$; they satisfy a system of ordinary differential equations, with respect to the variable x_1, with k and τ as parameters; this system can be written

$$\text{(3.30)}\qquad C\hat{u}_1 + \frac{\partial \hat{p}}{\partial x_1} = 0,$$

$$\text{(3.31)}\qquad C\hat{u}_2 + ik\hat{p} = 0,$$

$$\text{(3.32)}\qquad \frac{\partial \hat{u}_1}{\partial x_1} + ik\hat{u}_2 = 0.$$

If we eliminate the pressure from this system, by multiplying (3.30) by $-ik$, differentiating (3.31) with respect to x_1, and adding the two resulting inequalities, we obtain

$$\text{(3.33)}\qquad C\left(-ik\hat{u}_1 + \frac{\partial \hat{u}_2}{\partial x_1}\right) = 0.$$

With the help of (3.32), we eliminate $\hat{u}_2$; then $\hat{u}_1$ satisfies

$$C\left(\frac{\partial^2 \hat{u}_1}{\partial x_1^2} - |k|^2 \hat{u}_1\right) = 0.$$

This is an ordinary differential equation of the fourth order in x_1 parameterized by k and τ; its general solution is of the form

$$\alpha_1(k, \tau) \exp(-|k|x_1) + \beta_1(k, \tau) \exp(\lambda x_1) + \gamma_1(k, \tau) \exp(|k|x_1) + \delta_1(k, \tau) \exp(\lambda' x_1).$$

From Proposition 3.1, u_1 belongs to $L^2(\mathbb{R}_+ \times \mathbb{R} \times \mathbb{R})$ if $\mu > 0$, and so does $\hat{u}_1$; thus, for almost every k and τ, $\hat{u}_1(\cdot, k, \tau)$ is square integrable. Therefore, γ_1 and δ_1 vanish for almost every k and τ; thus,

$$\text{(3.34)}\qquad \hat{u}_1 = \alpha_1(k, \tau) \exp(-|k|x_1) + \beta_1(k, \tau) \exp(\lambda x_1).$$

Similarly, eliminating $\hat{u}_1$ from (3.33) and (3.34), we obtain

$$C\left(\frac{\partial^2 \hat{u}_2}{\partial x_1^2} - |k|^2 \hat{u}_2\right) = 0.$$

With the same argument as above,

$$\text{(3.35)}\qquad \hat{u}_2 = \alpha_2(k, \tau) \exp(-|k|x_1) + \beta_2(k, \tau) \exp(\lambda x_1).$$

The divergence-free relation (3.32) implies immediately that

$$\text{(3.36)}\qquad -|k|\alpha_1 + ik\alpha_2 = 0, \quad \lambda\beta_1 + ik\beta_2 = 0.$$

If we take into account the boundary conditions,

$$\hat{g}_1 = \alpha_1 + \beta_1, \quad \hat{g}_2 = \alpha_2 + \beta_2, \tag{3.37}$$

we will now express α_1, α_2, β_1, and β_2 in terms of $\hat{g}_1$ and $\hat{g}_2$. To obtain α_1 and β_1. we have to divide by $\lambda + |k|$, which vanishes only for $k = \tau = 0$. In order to obtain α_2 and β_2, we have to divide moreover by k. In order to obtain locally integrable functions α and β, we have to show that $\lambda/\lambda + |k|$ is bounded in a neighborhood of zero. But,

$$\lambda = -2\frac{i(\tau + a_2 k) + |k|^2}{a_1 + \rho}$$

and

$$\lambda + |k| = -2\frac{\tau + a_2 k}{a_1 + \rho + 2\nu|k|},$$

which shows immediately the desired estimate. Eliminating between (3.36) and (3.37), we can write

$$\alpha_1 = \frac{\lambda\hat{g}_1 + ik\hat{g}_2}{\lambda + |k|} = \frac{\lambda(\hat{g}_1 - i\sigma\hat{g}_2) + i\sigma(\lambda + |k|)}{\lambda + |k|},$$

and there is a locally integrable function α_1 which is almost everywhere equal to the expression given by the first of formulae (3.29). The second of these formulae gives a locally integrable β_1, because $\beta_1 = \hat{g}_1 - \alpha_1$. From the first formula of (3.36) divided by k, α_2 is locally integrable. From the second formula of (3.36) and the expression of β_1, a division by k gives the formula for β_2. If α' and β' are distributions which satisfy (3.28), they differ from a and β by distributions S and T with support in $\mathbb{R}_k \times \{0\}_\omega$. We must have

$$S \exp(-|k|x_1) + T \exp(\lambda x_1) = 0, \text{ for all } x_1.$$

This is possible only if S and T have their support in $\{0\}_{k,\omega}$, and $T + S = 0$. Therefore, we can say that α and β are respectively determined up to the addition of S and $-S$, with S a distribution with support in $\{0\}_{k,\omega}$. □

The most important consequence of the previous proposition is the following result on the operator, which assigns to the boundary value of u the normal constraint at the boundary.

COROLLARY 3.3. *Let (u, p) be the solution of* (3.9). *Let E be the two-by-two matrix given by*

$$E(k, \tau) = -\begin{pmatrix} (i\tau/|k|) + ia_2\sigma + \nu(|k| - \lambda) & i\sigma(\nu\lambda - a_1) \\ -i\sigma\nu\lambda & \nu(|k| - \lambda) \end{pmatrix}. \tag{3.38}$$

Then the normal constraint $(-p + \nu\,\partial u_1/\partial x_1,\ \nu\,\partial u_2/\partial x_1)|_\Sigma = \sigma_n$ is given by

$$\sigma_n = \mathscr{E}g,$$

or equivalently

$$\hat{\sigma}_n(k, \tau) = E(k, \tau)\hat{g}(k, \tau).$$

Proof. From (3.29), we compute $\partial\hat{u}_1/\partial x_1$ and $\partial\hat{u}_2/\partial x_1$ on the boundary:

$$\frac{\partial\hat{u}}{\partial x_1} = -|k|\alpha + \lambda\beta = -|k|\alpha + \lambda(\hat{g} - \alpha) = \lambda\hat{g} - \begin{pmatrix} \lambda\hat{g}_1 + ik\hat{g}_2 \\ -i\sigma(\lambda\hat{g}_1 + ik\hat{g}_2) \end{pmatrix}.$$

We observe that the distribution S does not contribute, because $-(|k|+\lambda)S=0$. The above formula proves that

$$\frac{\partial \hat{u}_1}{\partial x_1} = -ik\hat{g}_2, \tag{3.39}$$

$$\frac{\partial \hat{u}_2}{\partial x_1} = i\lambda\sigma\hat{g}_1 + (\lambda - |k|)\hat{g}_2. \tag{3.40}$$

To obtain an expression for the pressure, we observe that, from (3.31), we can write

$$ik\hat{p} = -C\hat{u}_2 = -P(k, \tau; -|k|)\alpha_2 \exp(-|k|x_1) - P(k, \tau; \lambda)\beta_2 \exp(\lambda x_1).$$

The second term of this expression vanishes by definition of λ; in the first term,

$$P(k, \tau; |k|) = -a_1|k| + i(\tau + a_2 k).$$

Thus,

$$ik\hat{p} = \frac{i\sigma(\lambda\hat{g}_1 + ik\hat{g}_2)(-a_1|k| + i(\tau + a_2 k))}{\lambda + |k|}.$$

This expression can be simplified by algebraic manipulations: let $y = \lambda + |k|$; then

$$P(k, \tau, y - |k|) = 0$$

so that

$$-\nu y^2 + 2\nu|k|y + a_1 y - a_1|k| + i(\tau + a_2 k) = 0.$$

Hence,

$$y(-\nu y + 2\nu|k| + a_1) = a_1|k| - i(\tau + a_2 k),$$

that is

$$\frac{a_1|k| - i(\tau + a_2 k)}{\lambda + |k|} = a_1 + \nu|k| - \nu\lambda.$$

If we substitute this in the above expression of $ik\hat{p}$, we obtain

$$ik\hat{p} = -i\sigma(a_1 + \nu|k| - \nu\lambda)(\lambda\hat{g}_1 + |k|\hat{g}_2).$$

Using once again the equation which defines λ, we have

$$a_1\lambda + \nu|k|\lambda - \nu\lambda^2 = \nu|k|\lambda - \nu|k|^2 - i(\tau + a_2 k),$$

and from here, we obtain

$$ik\hat{p} = ik\left[\hat{g}_1\left\{\nu(|k| - \lambda) + i\left(a_2\sigma + \frac{\tau}{|k|}\right)\right\} + \hat{g}_2 i\sigma\{\nu(\lambda - |k|) - a_1\}\right].$$

If we divide both sides of this equation by k, we obtain, with the help of (3.39) the unique locally integrable function equal to $\hat{p}$; thus

$$-\nu\frac{\partial \hat{u}_1}{\partial x_1} + \hat{p} = \hat{g}_1\left\{(\nu(|k| - \lambda) + i\left(a_2\sigma + \frac{\tau}{|k|}\right)\right\} + \hat{g}_2 i\sigma(\nu\lambda - a_1).$$

Putting together this last expression and (3.40), we obtain the matrix E. A distribution $\hat{p}$ which is solution of our equations is equal to the expression $\hat{g}_1\{\nu(|k| - \lambda) + i(a_2\sigma + |k|^{-1}\tau)\} + \hat{g}_2 i\sigma\{\nu(\lambda - |k|) - a_1\}$ up to the addition of a distribution with support in $\{k = 0\}$. This corresponds to the addition of a space independent constant to p,

which is natural because p is defined only via its gradient. But, as our problem was obtained through linearization, around a constant velocity field and a constant pressure, this constant is equal to zero. □

3.2. Variational formulation for the problem with transparent condition. To obtain a tractable formulation for the Oseen problem and "transparent" boundary conditions, we will write a variational formulation. The existence of a solution of the variational problem will be ensured by simply taking the restriction of a full-space problem. A functional problem has to be settled in order to justify this formulation; a certain pseudo-differential operator $\mathscr{L}$ acts on the trace of u on Σ. Originally $\mathscr{L}$ is defined only on very smooth functions, and it has to be extended to a larger functional class. After this is done, we have a clean presentation of the problem with "transparent" condition. The quotes will be removed only after uniqueness is proved.

Let μ be strictly positive, let f satisfy the assumptions of Lemma 2.9, and let u^0 equal zero. According to Lemma 2.9, the solution (u, p) of (2.29)-(2.31) belongs to $C^\infty(\mathbb{R}^+; H^\infty(\mathbb{R}^2))$. Assume that f satisfies (2.37). We assume moreover that:

$$f \in \mathscr{S}((0,\infty); \mathbf{H}^\infty(\mathbb{R}^2)). \tag{3.41}$$

In particular, f vanishes of infinite order at $t=0$. Then, the solution of (2.29)-(2.31) extended by 0 for $t \leqq 0$ belongs to $C^\infty(\mathbb{R}; \mathbf{H}^\infty(\mathbb{R}^2))$. Arguing as in Proposition 3.1, u belongs to $\mathscr{S}(\mathbb{R}; \mathbf{H}^\infty(\mathbb{R}^2))$, and its trace on Σ belongs to $\mathscr{S}(\mathbb{R}; \mathbf{H}^\infty(\mathbb{R}))$. In the region $\Omega_+ \times \mathbb{R}$, u satisfies (3.9), with

$$g = u|_\Sigma.$$

If we still denote u the restriction of u to $\Omega_- \times \mathbb{R}$, and if we multiply the equation satisfied by u on $\Omega_- \times \mathbb{R}$ by an arbitrary v in $W(\Omega_-)$, and integrate, we obtain, thanks to Green's formula

$$(u_t, v) + \mu(u, v) + \tilde{\mathbf{a}}(u, v) + \int_{\mathbb{R}} \left(\frac{a_1}{2} g \cdot v|_\Sigma - \sigma_n \cdot v|_\Sigma \right) dx_2 = (f, v).$$

From Corollary 3.3, this relation can be written

$$(u_t, v) + \mu(u, v) + \tilde{\mathbf{a}}(u, v) + \int_{\mathbb{R}} \left[\left\{ \frac{a_1}{2} g - \mathscr{E} g \right\} \cdot v|_\Sigma \right] dx_2 = (f, v). \tag{3.42}$$

Let us write down the matrix $a_1(I/2) + E$:

$$\frac{a_1 I}{2} - E(k, \tau) = \begin{pmatrix} (a_1/2) + (i\tau/|k|) + ia_2\sigma + \nu(|k| - \lambda) & i\sigma(\nu\lambda - a_1) \\ -i\sigma\nu\lambda & (a_1/2) + \nu|k| - \lambda \end{pmatrix}. \tag{3.43}$$

We can decompose this matrix into the sum of two matrices:

$$\frac{a_1 I}{2} - E = K_1 + L,$$

where K_1 is given by

$$K_1 = \begin{pmatrix} 1/|k| & 0 \\ 0 & 0 \end{pmatrix} i\tau. \tag{3.44}$$

Let $\mathscr{K}_1$ be the operator of symbol K_1. From the definition (2.18) of the scalar product s, we can see that

$$(u_t, v) + \mu(u, v) + \langle \mathscr{K}_1 u, v \rangle = s(u_t, v) + \mu s(u, v). \tag{3.45}$$

The matrix L can be written as follows, recalling that $(a_1/2)-\nu\lambda=\rho/2$:

$$L=\begin{pmatrix} ia_2\sigma+\nu|k|+\dfrac{\rho}{2} & \dfrac{-i\sigma(a_1+\rho)}{2} \\ \dfrac{i\sigma(\rho-a_1)}{2} & \nu|k|+\dfrac{\rho}{2} \end{pmatrix}. \tag{3.46}$$

It is convenient to write L as a sum of two matrices, one of which has ρ as a factor:

$$L=M+N_1,\ N_1=\rho\frac{N}{2}. \tag{3.47}$$

The expressions of M and N are given by

$$M=\begin{pmatrix} ia_2\sigma+\nu|k| & -i\sigma a_1/2 \\ -i\sigma a_1/2 & \nu|k| \end{pmatrix}, \tag{3.48}$$

$$N=\begin{pmatrix} 1 & -i\sigma \\ i\sigma & 1 \end{pmatrix}. \tag{3.49}$$

The matrix N is Hermitian positive semidefinite; we have

$$\mathcal{N}u=(u_1+\mathcal{H}u_2,\ \mathcal{H}(u_1+\mathcal{H}u_2)). \tag{3.50}$$

This relation explains why, as we mentioned in § 2, the trace of $u_1+\mathcal{H}u_2$ on Σ plays a particular role in the analysis of transparent boundary conditions. As $\mathrm{Re}(\rho)$ is strictly positive, we have

$$\mathrm{Re}\,(N\hat{u},\hat{u})=(\mathrm{Re}\,\rho)|\hat{u}_1-i\sigma\hat{u}_2|^2\geqq 0.$$

The matrix M is not Hermitian, but the symmetrized M, $(M+M^*)/2$ is equal to

$$\frac{M+M^*}{2}=\begin{pmatrix} \nu|k| & 0 \\ 0 & \nu|k| \end{pmatrix}. \tag{3.51}$$

This is a Hermitian positive definite matrix, for all nonzero k. Therefore L satisfies

$$\mathrm{Re}\,(L\hat{u},\hat{u})\geqq 0 \quad \forall\hat{u}\text{ in }\mathbb{C}^2. \tag{3.52}$$

In particular, $L+I$ is invertible, for all values of the parameters τ and k, and in Hermitian norm on $\mathbb{C}^2$, we have

$$|(L+I)^{-1}\hat{u}|\leqq|\hat{u}| \quad \forall\hat{u}\text{ in }\mathbb{C}^2. \tag{3.53}$$

We have thus shown the following result.

PROPOSITION 3.4. *For any f satisfying* (2.37) *and* (3.41), *for $u^0=0$, and for any positive μ, let (u,p) be the solution of* (2.29)-(2.31). *If we denote still by u the function u restricted to $\Omega_-\times\mathbb{R}^+$ and extended by 0 for $t\leqq 0$, then u satisfies the variational inequality*

$$s(u_t,v)+\mu s(u,v)+\tilde{\mathbf{a}}(u,v)+\langle\mathcal{L}u,v\rangle=(f,v) \quad \forall v\in W(\Omega_-), \tag{3.54}$$

where the pseudo-differential operator $\mathcal{L}$ is defined by its symbol L, given by (3.46).

3.3. Functional analysis of the operator $\mathcal{L}$. The class of data u^0 and f for which we have a solution of (3.54) is much too restricted. Thus, we extend it in several steps, by giving first a suitable definition of the domain of $\mathcal{L}$. We rely, of course, on the positivity of the matrix L observed in (3.47) to (3.52). Then, we shall give a dense subset of the domain of $\mathcal{L}$, and prove that $\mathcal{L}$ is a causal operator. Finally, the expression $\int_0^T\langle(\mathcal{L}u)(t),u(t)\rangle\,dt$ is greater than or equal to zero, if u is in the domain of $\mathcal{L}$, and vanishes for $t\leqq 0$.

The operator $\mathscr{L}$ is a pseudo-differential operator that belongs to the class $S^1_{1,0}$ of [17], [20], and [30]; it belongs in fact to a certain anisotropic class, which could be defined. As we do not seek the highest possible generality, we will be content with results specific to our variational problem:

DEFINITION 3.5. The domain $D(\mathscr{L})$ of $\mathscr{L}$ is the space of functions g such that there exists an h in $L^2(\mathbb{R};\ W^{-1/2}(\mathbb{R})\times H^{-1/2}(\mathbb{R}))$ such that

$$\hat{g} = (I+L)^{-1}\hat{h}. \tag{3.55}$$

Then, on $D(\mathscr{L})$, $\mathscr{L}$ is defined by

$$\mathscr{L}g = h - g. \tag{3.56}$$

An obvious consequence of the definition and of estimate (3.53) is that $D(\mathscr{L})$ is a subspace of $L^2(\mathbb{R};\ W^{1/2}\times H^{-1/2})$. A sufficient condition for g to belong to the domain of $\mathscr{L}$ is given by the following lemma.

LEMMA 3.6. *The set Y of functions g on Σ such that*

$$g\in L^2(\mathbb{R};\mathbf{H}^{1/2}(\mathbb{R})) \quad \textit{and} \quad g_1+\mathscr{H}g_2\in H^{1/2}(\mathbb{R};\ H^{-1/2}(\mathbb{R})) \tag{3.57}$$

is a subset of $D(\mathscr{L})$; moreover, if Z is the space defined at (3.11), *$\mathscr{S}(\mathbb{R};\ Z)$ is dense in $L^2(\mathbb{R};\ W^{1/2}\times H^{-1/2})$ in the following sense; for every g in $D(\mathscr{L})$, there exists a sequence of elements g_n of $\mathscr{S}(\mathbb{Q};\ Z)$ such that g_n converges to g and $\mathscr{L}g_n$ converges to $\mathscr{L}g$ in $L^2(\mathbb{R};\ W^{-1/2}\times H^{-1/2})$.*

Proof. According to estimate (3.27), there exists a decomposition

$$\rho = \rho_1+\rho_2,$$

such that

$$|\rho_1|\leqq \gamma\sqrt{1+|k|^2},\quad |\rho_2|\leqq \gamma\sqrt[4]{1+|\tau|^2}.$$

We decompose L as a sum

$$L=\left(M+\frac{\rho_1 N}{2}\right)+\left(\frac{\rho_2 N}{2}\right).$$

The first functional assumption on g implies that

$$\int_{\mathbb{R}}\int_{\mathbb{R}}\sqrt{1+|k|^2}|\hat{g}|^2\,dk\,d\omega<+\infty.$$

The coefficients of M are bounded by $\gamma'(1+|k|)$, and thus,

$$\int_{\mathbb{R}}\int_{\mathbb{R}}\frac{1}{\sqrt{1+|k|^2}}\left|\left(M+\frac{\rho_1 N}{2}\right)\hat{g}\right|^2 dk\,d\omega\leqq(\gamma+\gamma')\int_{\mathbb{R}}\int_{\mathbb{R}}\sqrt{1+|k|^2}|\hat{g}|^2\,dk\,d\omega<+\infty.$$

The second functional assumption means that

$$\int_{\mathbb{R}}\int_{\mathbb{R}}\sqrt{\frac{1+|\tau|^2}{1+|k|^2}}|\hat{g}_1-i\sigma\hat{g}_2|^2\,dk\,d\omega<+\infty.$$

The second piece of $L\hat{g}$ is estimated by

$$\int_{\mathbb{R}}\int_{\mathbb{R}}\frac{|\rho_2 N\hat{g}|^2}{\sqrt[4]{1+|k|^2}}\,dk\,d\omega\leqq\gamma\int_{\mathbb{R}}\int_{\mathbb{R}}\sqrt{\frac{1+|\tau|^2}{1+|k|^2}}|\hat{g}_1-i\sigma\hat{g}_2|^2\,dk\,d\omega<+\infty.$$

This proves that $\mathscr{L}g$ belongs to $L^2(\mathbb{R},\ W^{-1/2}(\mathbb{R})\times H^{-1/2}(\mathbb{R}))$, since $H^{-1/2}$ is a subspace of $W^{-1/2}$. On the other hand, g belongs to $L^2(\mathbb{R};\ W^{-1/2}(\mathbb{R})\times H^{-1/2}(\mathbb{R}))$, and thus, g belongs to $D(\mathscr{L})$.

To build a sequence g_n with the required properties, let g belong to $D(\mathscr{L})$, and let $h = \mathscr{L}g$; let φ_n, ψ_n and χ_n be elements of $\mathscr{S}(\mathbb{R})$ such that

$$\hat{\varphi}_n(k) = 0 \text{ if } |k| \leqq \frac{1}{n} \text{ or if } |k| \geqq n;\ \hat{\varphi}_n = 1 \quad \text{if } \frac{2}{n} \leqq |k| \leqq n-1,$$

$$0 \leqq \hat{\varphi}_n \leqq 1 \text{ everywhere;}$$

$$\psi_n(t) = n\psi(nt-1) \text{ with } \psi \text{ belonging to } \mathscr{D}(\mathbb{R}),\ \int_{\mathbb{R}} \psi\, dx = 1,$$

$$\chi_m(t) = \exp\left(-\frac{t^2}{2m}\right) \Rightarrow \hat{\chi}_m(\omega) = \sqrt{\frac{m}{2\pi}} \exp\left(-\frac{m\omega^2}{2}\right).$$

We define

$$h_n = h * (\varphi_n \psi_n).$$

Clearly, h_n belongs to $\mathbf{H}^\infty(\mathbb{R}, Z)$, because the choice of φ_n disposes of all problems at $k = 0$. Moreover, for any given ε, there exists an $n(\varepsilon)$ such that for $n > n(\varepsilon)$,

$$\|h - h_n\|_{L^2(\mathbb{R}; W^{-1/2}(\mathbb{R}) \times H^{-1/2}(\mathbb{R}))} \leqq \frac{\varepsilon}{2}.$$

Fix n such that this inequality holds. We define now

$$g_n = (\mathscr{L}+1)^{-1} h_n;\ g_{mn} = \chi_m g_n.$$

From estimate (3.53),

$$\|g - g_n\|_{L^2(\mathbb{R};\ W^{-1/2}(\mathbb{R}) \times H^{-1/2}(\mathbb{R}))} \leqq \frac{\varepsilon}{2}.$$

Let

$$h_{mn} = (\mathscr{L}+1) g_{mn},$$

which is well defined thanks to the first part of the lemma. In the Fourier variable,

$$\hat{h}_{mn}(k, \tau) - \hat{h}_n(k, \tau) = (L(k, \tau)+1)\{(\hat{\chi}_m(\cdot) - \delta^\omega) * \hat{g}_n(k, \cdot)\}(\tau).$$

As m tends to infinity, $\hat{\chi}_m * \hat{g}_n$ converges to $\hat{g}_n$ in $\mathscr{F}\mathbf{H}^m(\Sigma)$, for all m. As the multiplication by $L(k, \tau)$ maps $\mathscr{F}\mathbf{H}^1(\Sigma)$ continuously into $\mathscr{F}\mathbf{L}^2(\Sigma)$, there exists an m such that

$$\|h_{mn} - h_n\|_{L^2(\mathbb{R};\ W^{-1/2}(\mathbb{R}) \times H^{-1/2}(\mathbb{R}))} \leqq \frac{\varepsilon}{2}.$$

this proves the density of $\mathscr{S}(\mathbb{R}; Z)$ in $D(\mathscr{L})$. □

Now, we are able to prove that $\mathscr{L}$ is a causal operator.

LEMMA 3.7. *If g belongs to $D(\mathscr{L})$, and vanishes for $t < 0$, then $\mathscr{L}g$ vanishes for $t < 0$.*

Proof. Let g be an element of $D(\mathscr{L})$, and let $h = (\mathscr{L}+1)g$; assume that g vanishes for $t \leqq 0$. Let φ_n, ψ_n and g_n, h_n be as in the previous lemma. Then, as we can see in the Fourier variable,

$$g_n = g * (\varphi_n \psi_n).$$

The choice we made of ψ_n implies that

$$g_n = 0 \text{ for } t > 0.$$

The choice of φ_n implies that

$$\hat{g}_n(k, \omega) = 0 \text{ for } |k| \geqq n.$$

If θ is a square integrable function on $\mathbb{R}$ that vanishes on $\mathbb{R}_-$, we can write

$$\hat{\theta}(\omega - i\mu) = \int_0^\infty \theta(t) \exp(-it(\omega - i\mu))\, dt,$$

for $\mu = 0$, with the estimate

$$|\hat{\theta}(\omega - i\mu)| \leqq |\theta|_{L^2(\mathbb{R})} \sqrt{\int_0^\infty e^{-2t\mu}\, dt} \leqq |\theta|_{L^2(\mathbb{R})} \frac{1}{\sqrt{2\mu}}.$$

Therefore, for g_n, we have the estimate

$$|\hat{g}_n(k, \omega - i\mu)| \leqq |\mathcal{F}_{x_2 \to k} g_n(k, \cdot)|_{L^2(\mathbb{R}^t)} \frac{1}{\sqrt{2\mu_0}} \text{ if } \mu \geqq \mu_0 > 0,$$

where $L^2(\mathbb{R}^t)$ is the space of square integrable functions of time.
We define an operator $\mathcal{L}_k$ by its symbol:

$$\mathcal{L}_k v = (\mathcal{F}_{t \to \omega})^{-1}\{L(k, \cdot)\tilde{v}(\cdot)\},$$

where the tilda $\tilde{}$ denotes the Fourier transform with respect to the time variable. If v belongs to $L^2(\mathbb{R})$, then $\mathcal{L}_k v$ belongs to $H^{-1/2}(\mathbb{R})$. The root ρ can be extended to the half-plane $\tau = \omega - i\mu$, $\mu \geqq 0$, as an analytic function of ω, with the estimate

$$|\rho(k, \omega)| \leqq C(1 + |k| + \sqrt{|\omega|} + \sqrt{|\mu|}).$$

We can apply the Paley-Wiener-Schwarz theorem to $\mathcal{L}_k \mathcal{F}_{x_2 \to k} g_n(k, \cdot)$; from the previous considerations, for almost every k, $L(k, \tau)\hat{g}_n(k, \tau)$ is analytic in τ for $\mu = -\mathrm{Im}(\tau) > 0$, and satisfies an estimate of the form

$$|L(k, \tau)\hat{g}_n(k, \tau)| \leqq C(k, \mu_0)(1 + |\tau| + |\mu|) \quad \forall \mu \geqq \mu_0.$$

Thus, for almost every k, $\omega \to L(k, \omega - i\mu)\hat{g}_n(k, \omega - i\mu)$ is the Fourier-Laplace transform of a distribution with support in $[0, \infty)$. On the other hand, $L(k, \omega - i\mu)\hat{g}_n(k, \omega - i\mu)$ is square integrable with respect to ω, and therefore,

$$(\mathcal{F}_{t \to \omega})^{-1}(L(k, \cdot)\hat{g}_n(k, \cdot)) = 0 \quad \forall\, t \leqq 0, \text{ and a.e. } k,$$

and finally,

$$\mathcal{L} g_n = 0 \quad \forall\, t \leqq 0, \text{ and for all } x_2.$$

By density, the same result holds for g. □

An immediate consequence of Lemma 3.7 is the following.

LEMMA 3.8. *Let u^0 belong to $\mathbf{H}(\mathbb{R}^2)$, and f to $L^2(0, \infty; \mathbf{L}^2(\mathbb{R}^2))$. Assume that u^0 and f satisfy respectively the support conditions* (2.36) *and* (2.37). *If μ is strictly positive, and if (u, p) is the solution of* (2.29)-(2.31), *extended by zero for negative time, then the trace of u on Σ belongs to $D(\mathcal{L})$.*

Proof. According to Proposition 2.7, u belongs to $L^2(0, \infty, W(\Omega_-))$, and according to Proposition 2.4, the trace of u on Σ belongs to $L^2(0, \infty, W^{1/2} \times H^{1/2})$. On the other hand, the support condition enables us to apply Lemma 2.14, and therefore, $(u_1 + \mathcal{H}u_2)|_\Sigma$ belongs to $H^\infty(\Sigma)$. □

We first obtain a result of existence, under the support condition on u^0 and f, with some functional analysis on $\mathcal{L}$.

PROPOSITION 3.9. *For all u^0 in $H(\Omega_-)$ and all f in $L^2(0,\infty,L^2(\Omega_-))$ that satisfy the support conditions* (2.36), (2.37), *and for all strictly positive μ, there exists a function u satisfying the variational equality* (3.54) *with initial data u^0, and such that*

$$u \in L^\infty([0,\infty);\, H(\Omega_-)),$$
$$\nabla u \in L^2([0,\infty);\, L^2(\Omega_-)),$$
$$u_t \in L^2([0,\infty);\, W'(\Omega_-)).$$

Proof. We have to approximate a right-hand side $f + u^0 \otimes \delta'$ by a smooth right-hand side f_n belonging to $\mathscr{S}([0,\infty);\, H^\infty(\mathbb{R}^2))$. This is clearly possible by truncation and regularization. The passage to the limit in the variational inequality is easy.

We will obtain a much better result of existence, not only for the sake of exhaustivity, but also to have a convenient frame for the proof of uniqueness. We write (3.52) as a problem with time in the full real line, extending u by 0 for $t \geqq 0$; after an integration by parts in time, we have for all v in $L^2(\mathbb{R};\, W(\Omega_-))$ such that v_t belongs to $L^2(\mathbb{R};\, H(\Omega_-))$

$$(3.58)\quad \int_{\mathbb{R}} \{-s(u, v_t) + \mu s(u, v) + \tilde{\mathbf{a}}(u, v)\}\, dt + \langle u, \mathscr{L}^* v\rangle_\Sigma = s(u^0, v(0)) + \int_{\mathbb{R}} (f, v)\, dt.$$

Here, of course, $\mathscr{L}^*$ has L^* for symbol, the domain of $\mathscr{L}^*$ is defined similarly to the domain of $\mathscr{L}$, and the trace of a test function v on Σ belongs to $D(\mathscr{L}^*)$ thanks to Lemma 3.6 and simple interpolation. We need to define three more operators to prove existence and uniqueness.

DEFINITION 3.10. The operator A is an operator from $W(\Omega_-)$ to $W'(\Omega_-)$ defined by

$$(3.59)\qquad s(Au, v) = \tilde{\mathbf{a}}(u, v) \quad \forall v \in W(\Omega_-).$$

The operator $B(\tau)$ is an operator from $W(\Omega_-)$ to $W'(\Omega_-)$ defined by

$$(3.60)\qquad s(B(\tau)u, v) = \int_{\mathbb{R}} L(k, \tau)\hat{u}(0, k)\bar{\hat{v}}(k, \tau)\, dk, \quad \forall v \in W(\Omega_-).$$

The operator S is an operator from $L^2(\Omega_-)$ to $H(\Omega_-)$ defined by

$$(3.61)\qquad s(Sf, v) = (f, v) \quad \forall v \in H(\Omega_-).$$

The operator A is well defined; this is a classical result. The mapping which assigns to a pair $(u; v)$ belonging to $W(\Omega_-) \times W(\Omega_-)$ the expression

$$\int_{\mathbb{R}} L(k, \tau)\hat{u}(0, k)\bar{\hat{v}}(0, k)\, dk$$

is clearly a sesquilinear continuous mapping, and thus $B(\tau)$ is well defined. Finally, the mapping

$$v \to (f, v)$$

is linear continuous on $H(\Omega_-)$, and Sf is thus well defined.

The existence and uniqueness theorem reads.

THEOREM 3.11. *For all u^0 in $H(\Omega_-)$ and all f in $L^2(0,\infty;\mathbf{L}^2(\Omega_-))$, and for all strictly positive μ, there exists a unique u in $L^\infty([0,\infty);\, H(\Omega_-)) \cap L^2([0,\infty);\, W(\Omega_-))$ such that u_t belongs to $L^2([0,\infty);\, W(\Omega_-))$ and*

$$(3.62)\qquad \int_{\mathbb{R}} \{-s(u, v_t) + \mu s(u, v) + \tilde{\mathbf{a}}(u, v)\}\, dt + \langle u, \mathscr{L}^* v\rangle_\Sigma = s(u^0, v(0)) + \int_{\mathbb{R}} (f, v)\, dt.$$

for all v in $L^2(\mathbb{R};\, W(\Omega_-))$ such that v_t belongs to $L^2(\mathbb{R};\, H(\Omega_-))$.

In order to prove this theorem, we need a positivity result which will be useful for the existence part.

LEMMA 3.12. *Let g belong to the space Y defined at* (3.57). *Assume that u vanishes for negative time. Then, for all positive T, we have*

$$\int_0^T (\mathscr{L}g, g)\, dx_2\, dt \geqq 0.$$

Proof. Assume first that g belongs to $\mathscr{S}(\mathbb{R}; Z)$, and vanishes for negative time. Let u be the solution of (3.9). Then, if we multiply the equation $\mathscr{A}_\mu(u, p) = 0$ by u and integrate over Ω_+, we obtain

$$\int_0^T (\mathscr{L}g, g)\, dt = \|u(\cdot, T)\|^2 + \mu \int_0^T \{\tilde{\mathbf{a}}(u, u) + \|u(\cdot, t)\|^2\}\, dt - (\mathscr{K}_1 g_1(\cdot, T), g_1(\cdot, T))_\Gamma$$
$$- \mu \int_0^T (\mathscr{K}_1 g_1(\cdot, t), g_1(\cdot, t))_\Gamma\, dt,$$

where $\mathscr{K}$ is the boundary operator of symbol $|k|^{-1}$. We check that $\|u\|^2 - (\mathscr{K}g, g)_\Gamma$ is greater than or equal to zero. By an argument analogous to the one we used at proposition 2.4, we have the identity

$$(\mathscr{K}g_1, g_1)_\Gamma = -2(u_1, \mathscr{K}u_2),$$

and

$$\|u\|^2 - (\mathscr{K}u_1, u_1) = \|u_1 - \mathscr{K}u_2\|^2 \geqq 0.$$

Thus the positivity holds for smooth data. By density and causality, it will hold as stated in the statement of the lemma. □

Proof of the theorem. We first prove uniqueness; denote $\tilde{u}$ the partial Fourier transform of u in time. If u is the solution of (3.62), it satisfies

$$i\omega\tilde{u}(\tau) + \mu\tilde{u}(\tau) + A\tilde{u}(\tau) + B(\tau)\tilde{u}(\tau) = u^0 + S\tilde{f}(\tau). \tag{3.63}$$

If the data vanish, we have

$$i\omega\tilde{u}(\tau) + \mu\tilde{u}(\tau) + A\tilde{u}(\tau) + B(\tau)\tilde{u}(\tau) = 0.$$

For almost every ω, $\tilde{u}(\tau)$ belongs to $W(\Omega_-)$. If we multiply the above equation scalarly by $\tilde{u}(\tau)$ and take the real part, we obtain

$$\operatorname{Re}\{\mu s(\tilde{u}(\tau), \bar{\tilde{u}}(\tau)) + \tilde{\mathbf{a}}(\tilde{u}(\tau), \bar{\tilde{u}}(\tau)) + s(B(\tau)\tilde{u}(\tau), \bar{\tilde{u}}(\tau))\} = 0. \tag{3.64}$$

The definition of $\tilde{\mathbf{a}}$ and $B(\tau)$ implies that the corresponding terms in (3.64) are nonnegative; there remains

$$\operatorname{Re}(\mu s(\tilde{u}(\tau), \bar{\tilde{u}}(\tau))) \leqq 0,$$

and this implies that $\tilde{u}$ vanishes almost everywhere. We can conclude the uniqueness of the solution of (3.62).

Let us prove the existence under the assumptions of our theorem; if u and f satisfy our support condition, we know that there exists a unique solution to (3.62). If this support condition is not satisfied, extend u^0 and f by 0 in Ω_+, and approximate these extended u^0 and f by translated data u_n^0 and f_n that satisfy the support condition. For the solution u_n the positivity of $\mathscr{L}$ implies the estimate

$$s(u_n(T), u_n(T)) + \int_0^T \{\mu s(u_n(t), u_n(t)) + \tilde{\mathbf{a}}(u_n(t), u_n(t))\}\, dt \leqq s(u_n^0, u_n^0) \leqq s(u^0, u^0).$$

An easy passage to the limit gives the result. □

4. Absorbing boundary conditions. This section is dedicated to the approximation of the pseudo-differential operator $\mathscr{L}$ by more manageable operators. We have seen that the symbol L of $\mathscr{L}$ is a two-by-two matrix which is algebraic in τ, k and $\sigma = \operatorname{sgn}(k)$. We would like our approximation of $\mathscr{L}$ to be local in space and time, which would mean that its symbol is polynomial or rational in τ and k. Unfortunately, we do not know how to approximate σ or $|k|$ by rational fractions of low degree; thus, we will be content with an approximation which is rational in τ, k and σ. The idea is to approximate the root ρ by a sequence of r_n which keeps the essential property $\operatorname{Re}(r_n) \geqq 0$. We need a number of technical results that precisely describes the properties of the sequence of approximations. These properties enable us to prove a rather weak existence and uniqueness theorem for the problem with absorbing boundary conditions. Then, we estimate the difference between the solution of the problem with absorbing boundary conditions and the solution of the problem with transparent conditions. From this estimate, we deduce a better existence theorem. If ν is small, the difference between the two solutions is small with a power of ν; when u^0 and f satisfy the support conditions (2.36), (2.37), the difference is estimated in a space of smooth functions.

4.1. Approximation of the symbol L. We approximate the transparent boundary conditions in constraint formulation. If we approximated the symbol of the operator $g \to (\partial u/\partial x_1, p)$, we could run into trouble and obtain an ill-posed problem.

We recall that we obtained at (3.47) a decomposition of the matrix $L(k, \tau)$, which we write now with an explicit dependance on the viscosity ν.

$$L(k, \tau, \nu) = M(k, \nu) + \rho(k, \tau, \nu)\frac{N(k)}{2}. \tag{4.1}$$

The matrix M is of degree 1 in ν; $N(k)$ and the Hermitian symmetrization of $M(k, \nu)$ are both positive semidefinite matrices.

Now, we have to approximate $\rho(k, \tau, \nu)$ so that the successive approximations, denoted r_n, will satisfy the essential property

$$\operatorname{Re}(r_n(k, \tau, \nu)) \geqq 0 \quad \forall k, \omega, \nu. \tag{4.2}$$

This problem has been solved in [13] but with very few proofs. The principle of the approximation has been obtained by observing that for all complex number d we have

$$\lambda = \frac{\lambda d - |k|^2 - i\omega'\alpha}{-\lambda + \alpha + d}, \tag{4.3}$$

where $\alpha = a_1/\nu$ and $\omega' = (\tau + a_2 k)/a_1$.

The choice of $d = i\omega'$ and of the initialization $\lambda_1 = -i\omega'$ (because λ_0 might be a special case) defines a sequence recursively by:

$$\lambda_{n+1} = \frac{\lambda_n d - |k|^2 - i\omega'\alpha}{-\lambda_n + \alpha + d},$$

which has the three important properties:

(i) Each of the λ_n is rational in k, τ and σ;

(ii) The associated initial boundary value problem is, at least formally, well-posed because $\operatorname{Re}(\lambda_n) \leqq 0$, and we will prove that it is actually well-posed;

(iii) λ_n is the $[n-1, n-1]$ Padé approximant of λ around $\nu = 0$ (see [13]).

More precisely, we have the

DEFINITION 4.1. Denote

$$\omega' = \frac{\tau + a_2 k}{a_1}, \tag{4.4}$$

$$\alpha = \frac{a_1}{\nu}. \tag{4.5}$$

We define:

$$\lambda_0 = 0, \tag{4.6}$$

$$\lambda_1 = -i\omega', \tag{4.7}$$

$$\lambda_{n+1} = \frac{-i\omega'\lambda_n + i\alpha\omega' + k^2}{\lambda_n - \alpha - i\omega'}, \tag{4.8}$$

$$r_n = a_1 - 2\nu\lambda_n. \tag{4.9}$$

We will give now a sequence of technical lemmas on r_n and λ_n.

LEMMA 4.2. *For all* $n \geqq 0$, $k \in \mathbb{R}$, $\omega \in \mathbb{R}$, $\nu > 0$, r_n *is well defined and*

$$\operatorname{Re}(\lambda_n) \leqq 0. \tag{4.10}$$

Proof. For $n = 0$ or 1, the result is obvious. Assume that $\operatorname{Re}(\lambda_n) \leqq 0$; then, the expression $\lambda_n - \alpha + i\omega'$ has a strictly negative real part; moreover,

$$\lambda_{n+1} = \frac{(-i\omega'\lambda_n + i\alpha\omega' + k^2)(\bar{\lambda}_n - \alpha + i\omega')}{|\lambda_n - \alpha + i\omega'|^2}$$

so that,

$$\operatorname{Re}(\lambda_{n+1}) = \frac{[k^2 + \omega'^2] \operatorname{Re}(\lambda_n - \alpha)}{|\lambda_n - \alpha + i\omega'|^2},$$

and by induction, the results holds true. □

LEMMA 4.3. *Let* ρ *be as in* (3.24) *and* $\lambda = (a_1 - \rho)/2\nu$; *then, the following relation holds*:

$$\lambda_{n+1} - \lambda = -\frac{(\lambda_n - \lambda)[k^2 + \omega'^2]}{[\lambda_n - \alpha - i\omega'][\lambda - \alpha - i\omega']}. \tag{4.11}$$

Proof. We substitute $d = -i\omega'$ in (4.2)

$$\lambda = \frac{-i\omega'\lambda + i\alpha\omega' + k^2}{\lambda - \alpha - i\omega'}.$$

Therefore, subtracting this expression λ from the expression (4.7), we obtain immediately (4.10). □

The recursive definition of λ_n by a sequence of homographic transformations shows that λ_n is a rational fraction in k and ω. To obtain more information on the sequence λ_n, we study the particular case $\alpha = 1$. Then (4.8) becomes

$$\lambda_{n+1} = \frac{-i\omega'\lambda_n + i\omega' + k^2}{\lambda_n - i\omega' - 1}; \tag{4.12}$$

we define P_n and Q_n by

$$P_1 = -i\omega' \tag{4.13}$$

$$Q_1 = 1 \tag{4.14}$$

and

$$P_{n+1} = -i\omega' P_n + (i\omega' + k^2) Q_n \tag{4.15}$$

$$Q_{n+1} = P_n - (i\omega' + 1) Q_n. \tag{4.16}$$

Then, we have

$$\lambda_n = \frac{P_n}{Q_n}. \tag{4.17}$$

An obvious induction shows that P_n is at most of degree n, and that Q_n is at most of degree $n-1$, globally in k and ω'. Let P_n and Q_n be decomposed as a sum of globally homogeneous polynomials in k and ω', of decreasing degree:

$$P_n = P_n^n + P_n^{n-1} + \cdots + P_n^0, \tag{4.18}$$

$$Q_n = Q_n^{n-1} + Q_n^{n-2} + \cdots + Q_n^0; \tag{4.19}$$

we now define the polynomial Z_n by

$$Z_n = P_n + kQ_n. \tag{4.20}$$

Let

$$z = k - i\omega'. \tag{4.21}$$

Then Z_n satisfies the following relation:

$$Z_{n+1} = z(Z_n - Q_n), \tag{4.22}$$

and if we decompose Z_n into a sum of homogeneous polynomials

$$Z_n = Z_n^n + Z_n^{n-1} + \cdots + Z_n^1, \tag{4.23}$$

then we can deduce Q_n and P_n from Z_n by

$$P_n = i \operatorname{Im}(Z_n^1) + \operatorname{Re}(Z_n^2) + i \operatorname{Im}(Z_n^3) + \cdots \tag{4.24}$$

$$Q_n = \frac{\operatorname{Re}(Z_n^1) + i \operatorname{Im}(Z_n^2) + \operatorname{Re}(Z_n^3) + \cdots}{k}. \tag{4.25}$$

Of course, these expressions terminate differently according to the parity of n. □

LEMMA 4.4. *Assume that $\alpha = 1$. Then, there exists a polynomial R_n of degree $n-1$ in one variable such that Q_n^{n-1} can be written as*

$$Q_n^{n-1}(k, \omega') = k^{n-1} R_n\left(\frac{\omega'}{k}\right).$$

The zeros ζ_p of R_n are real and simple and except for $k = 0$, the next term $Q_n^{n-2}(k, k\zeta_p)$ does not vanish.

Proof. A simple induction shows that P_n^n and Q_n^{n-1} are given for n odd by

$$\begin{cases} P_n^n = i \operatorname{Im}(z^n), \\ Q_n^{n-1} = \operatorname{Re}(z^n)/k. \end{cases} \tag{4.26}$$

When n is even, they are given by

$$\begin{cases} P_n^n = \operatorname{Re}(z^n), \\ Q_n^{n-1} = [i \operatorname{Im}(z^n)]/k. \end{cases} \tag{4.27}$$

The polynomial $R_n(X)$ is given by

$$R_n(X) = i \operatorname{Im}(1 - iX)^i \quad \text{if } n \text{ is even},$$

$$R_n(X) = \operatorname{Re}(1 - iX)^n \quad \text{if } n \text{ is odd}.$$

From these formulae, R_n is of degree $n-1$ and the roots of R_n are real and simple. Geometrically, these roots are the ordinates of the intersection of the straight lines $\zeta = 1 + r e^{i\pi k/n}$ with the imaginary axis, when n is even; when n is odd, they are the abscissae of the intersection of these same lines with the real axis. There are exactly n of these lines. When n is even, one of these lines is parallel to the imaginary axis and another is the real axis; this yields the announced $n-1$ real solutions, one of which is zero. If n is odd, one of the lines is the real axis, which we discard, and there remain $n-1$ real distinct nonzero solutions. For the last assertion, we consider first the case of n even:

$$Q_n^{n-2} = \frac{\operatorname{Re}(Z_n^{n-1})}{k}.$$

Relation (4.22) implies that

$$Z_n = z^n - \sum_{j=1}^{n-1} z^j Q_{n-j}.$$

With the help of (4.26) and (4.25), we obtain

$$\begin{aligned}
\operatorname{Re}(Z_n^{n-1}) &= -\operatorname{Re}\left(\sum_{j=1}^{n-1} z^j Q_{n-j}^{n-j-1}\right) \\
&= -\operatorname{Re}\left(\sum_{\substack{j=1 \\ j \text{ even}}}^{n-1} z^j \frac{i \operatorname{Im} z^{n-j}}{k} + \sum_{\substack{j=1 \\ j \text{ odd}}}^{n-1} z^j \frac{\operatorname{Re} z^{n-j}}{k}\right) \\
&= \frac{1}{k}\left(\sum_{\substack{j=1 \\ j \text{ even}}}^{n-1} \operatorname{Im}(z^j) \operatorname{Im}(z^{n-j}) - \sum_{\substack{j=1 \\ j \text{ odd}}}^{n-1} \operatorname{Re}(z^j) \operatorname{Re}(z^{n-j})\right).
\end{aligned}$$

Assume now that z is a root of Q_n^{n-1} which does not vanish. Then, it is real and the above formula reduces to

$$\operatorname{Re}(Z_n^{n-1}) = -\frac{n}{2}\frac{z^n}{k},$$

which does not vanish.

The proof in the case of n odd is analogous and left to the reader. □

In the next lemma we give an estimate on $\lambda_n + (i\omega'/n)$, which will enable us to work on the variational problem for absorbing boundary conditions.

LEMMA 4.5. Assume that $\alpha = 1$. Then, for each n, there is a constant C_n, which depends neither on k nor on ω', such that

$$\left|\lambda_n + \frac{i\omega'}{n}\right| \leqq C_n(1+|k|)^2. \tag{4.28}$$

Proof. We observe that in

$$\lambda_n + \frac{i\omega'}{n} = \frac{nP_n + i\omega' Q_n}{nQ_n};$$

the term of highest degree in ω' of P_n is $(-i\omega')^n$, and the term of highest degree in ω' of Q_n is $(-i\omega')^{n-1}$, so that $nP_n + i\omega' Q_n$ does not have a term of degree n in ω'. Thus, we have the estimate

$$|nP_n + i\omega' Q_n| \leqq C_{1n}\{1 + |z|^{n-1} + |k||z|^{n-1}\}. \tag{4.29}$$

To estimate from below the denominator Q_n, we consider first the case when n is even. Let $(\zeta_p)_{1\leqq p\leqq n-1}$ denote the zeros of R_n, and let ξ_q denote the zeros of $X \to (Q_n^{n-2})(1, X)$. We make the convention that

$$\zeta_1 = 0.$$

Lemma 4.4 implies that there exists a β such that

$$\min_{\substack{1\leqq q\leqq n-2\\ 1\leqq p\leqq n-1}} |\xi_q - \zeta_p| = \beta > 0.$$

Therefore, this suggests an estimate on three different regions:

$$\mathcal{R}_1\colon \{k, \omega')/|k|^2+|\omega'|^2 < r^2\},$$

$$\mathcal{R}_2\colon \{(k, \omega')/|k|^2+|\omega'|^2 \geqq r^2 \text{ and } \min_p |\zeta_p - \omega'/k| \leqq \gamma\},$$

$$\mathcal{R}_3\colon \{(k, \omega')/|k|^2+|\omega'|^2 \geqq r^2 \text{ and } \min_p |\zeta_p - \omega'/k| > \gamma\}.$$

For any positive r, Q_n is bounded away from zero on $\mathcal{R}_1$; this is a consequence of Lemma 4.2. The precise choice of r and γ will be made below.

We can see that Q_n^{n-1} is the product with k^{n-1} of an imaginary polynomial R_n whose roots are all real. On the other hand, Q_n^{n-2} is real. Therefore,

$$|Q_n| \geqq \min(|Q_n^{n-1}|, |Q_n^{n-2}|) - \sum_{j=1}^{n-3} |Q_n^j|$$

$$\geqq \min(|Q_n^{n-1}|, |Q_n^{n-2}|) - \sum_{j=1}^{n-3}) C_j |z|^j,$$

where the C_j are positive numbers depending only on n. We have the relations

$$Q_n^{n-2} = K' k^{n-2} \prod_{q=1}^{n-2} \left(\frac{\omega'}{k} - \xi_q\right)$$

and

$$Q_n^{n-1} = K k^{n-1} \prod_{p=1}^{n-1} \left(\frac{\omega'}{k} - \zeta_p\right).$$

In the region $\mathcal{R}_2$, $|(\omega'/k) - \xi_p| \geqq \beta - \gamma$. Therefore, if we choose γ so that

$$\gamma \leqq \frac{\beta}{2}, \tag{4.30}$$

we obtain the estimate

$$|Q_n^{n-2}| \geqq |K'||k|^{n-2} \left(\frac{\beta}{2}\right)^{n-2}.$$

In the region $\mathcal{R}_2$, $|\omega'| \leqq |k|(\gamma + \max_p |\zeta_p|)$, and therefore, there exists a constant C_{2n} such that

$$|Q_n^{n-2}| \geqq C_{2n} |z|^{n-2}. \tag{4.31}$$

We argue similarly in the region $\mathcal{R}_3$; the homogeneous term Q_n^{n-1} satisfies

$$|Q_n^{n-1}| \geqq |K||k|^{n-1} \prod_{p=1}^{n-1} \left|\frac{\omega'}{k} - \zeta_p\right| = |K||\omega'|^{n-1} \prod_{p=1}^{n-1} \left|\frac{k\zeta_p}{\omega'} - 1\right|.$$

If $|k/\omega'| \leqq \varepsilon$, where ε is so chosen that $\varepsilon \max_p |\zeta_p| \leqq \frac{1}{2}$, then

$$|Q_n^{n-1}| \geqq |K||\omega'|^{n-1} \frac{1}{2^{n-1}\varepsilon^{n-1}} \geqq C_{3n}|z|^{n-1}.$$

If the converse inequality holds, then

$$|Q_n^{n-1}| \geqq |K||\gamma k|^{n-1} \geqq C_{4n}|z|^{n-1}.$$

Finally there exists a constant C_{5n} such that

$$|Q_n^{n-1}| \geqq C_{5n}|z|^{n-1} \quad \forall\, z \text{ in } \mathcal{R}_3. \tag{4.32}$$

The number r must be chosen so that

$$|z| \geqq r \Rightarrow \min\left(C_{2n}|z|^{n-2}, C_{5n}|z|^{n-1}\right) \geqq 2 \sum_{j=1}^{n-3} C_j |z|^j.$$

Then, there exists a constant C_{6n} such that

$$|Q_n| \geqq C_{6n}(1+|z|^{n-2}) \qquad \text{on } \mathcal{R}_2,$$

$$|Q_n| \geqq C_{6n}(1+|z|^{n-1}) \qquad \text{on } \mathcal{R}_3.$$

Finally,

$$\left|\lambda_n + \frac{i\omega'}{n}\right| \leqq \frac{C_{1n}\{1+|z|^{n-1}+|k||z|^{n-1}\}}{C_{6n}(1+|z|^{n-2})} \leqq C_{8n}(1+|z|+|k||z|) \text{ on } \mathcal{R}_2;$$

the inequality of the statement of the lemma is satisfied, because $|\omega'| \leqq$ constant $|k|$ on $\mathcal{R}_2$. On the other region,

$$\left|\lambda_n + \frac{i\omega'}{n}\right| \leqq \frac{C_{1n}(1+|z|^{n-1}+|k||z|^{n-1})}{C_{7n}(1+|z|^{n-1})} \leqq C_{9n}(1+|k|) \text{ on } \mathcal{R}_3.$$

Here, the inequality stated is clearly satisfied. No difficulty can come from region $\mathcal{R}_3$. The proof when n is odd is analogous and left to the reader. □

We can now state an estimate in the general case.

PROPOSITION 4.6. *There exists a constant c_n depending only on a_1 and a_2, such that*

$$\left|\lambda_n + \frac{i\omega'}{n}\right| \leqq c_n \frac{(1+\nu|k|)^2}{\nu}. \tag{4.33}$$

Proof. In the general case,

$$\lambda_{n+1} = \frac{-i\omega'\lambda_n + i\alpha\omega' + k^2}{\lambda_n - \alpha - i\omega'},$$

then,

$$\lambda_n(k, \omega', \alpha) = \frac{P_n(k, \omega', \alpha)}{Q_n(k, \omega', \alpha)},$$

where

$$P_n(k, \omega', \alpha) = P_n^n + \alpha P_n^{n-1} + \cdots + \alpha^n P_n^0,$$

$$Q_n(k, \omega', \alpha) = Q_n^{n-1} + \alpha Q_n^{n-2} + \cdots + \alpha^{n-1} Q_n^0,$$

or, in other terms,

$$P_n(k, \omega', \alpha) = \alpha^n P_n\left(\frac{k}{\alpha}, \frac{\omega'}{\alpha}, 1\right),$$

$$Q_n(k, \omega', \alpha) = \alpha^{n-1} Q_n\left(\frac{k}{\alpha}, \frac{\omega'}{\alpha}, 1\right).$$

Now, we can use (4.26):

$$\left|\lambda_n(k, \omega', \alpha) + \frac{i\omega'}{n}\right| = \left|\alpha\left\{\lambda_n\left(\frac{k}{\alpha}, \frac{\omega'}{\alpha}, 1\right) + \frac{i\omega'}{\alpha n}\right\}\right| \leqq C_n \alpha (1 + |k/\alpha|)^2.$$

This proves the proposition. □

4.2. Variational formulation: existence, uniqueness, error estimates. In this section, we shall state a variational formulation which is suitable for the analysis of absorbing conditions. Given data u^0 and f, we obtain by the Fourier method a unique solution of the problem with absorbing boundary conditions of any order n. This is not enough to obtain existence in nice spaces. One expects that the larger n, the closer the solution with artificial boundary conditions to the full space solution. This result is obtained by the analysis of the error, if the data satisfy the support conditions (2.36) and (2.37). It turns out that the error is in a better space than the solution u_n; using causality and positivity, we obtain the existence and uniqueness in convenient spaces.

DEFINITION 4.7. Let, for $n = 0$

$$L_0(k, \tau, \nu) = M(k, \tau, 0) + \frac{a_1}{2} N(k) \tag{4.34}$$

and

$$L_n(k, \tau, \nu) = M(k, \tau, \nu) + r_n(k, \tau, \nu) N(k)/2. \tag{4.35}$$

The operator $\mathscr{L}_n$ is defined by its symbol L_n:

$$\mathscr{L}_n(u) = \mathscr{F}^{-1}(L_n(\cdot, \cdot, \nu)\hat{u}(0, \cdot, \cdot)). \tag{4.36}$$

The variational formulation (3.58) is approximated by:

$$\int_{\mathbb{R}} \{-s(u, v_t) + \mu s(u, v) + \tilde{\mathbf{a}}(u, v)\}\, dt + \langle u, \mathscr{L}_n^* v\rangle_\Sigma = s(u^0, v(0)) + \int_{\mathbb{R}} (f, v)\, dt,$$
$$\forall\, v \in L^2(\mathbb{R}; W(\Omega_-)) \text{ such that } v_t \in L^2(\mathbb{R}; H(\Omega_-)). \tag{4.37}$$

We have a first result of existence and uniqueness, as follows.

PROPOSITION 4.8. *Let u^0 belong to H, and let f belong to $L^2(\mathbb{R}; H(\Omega_-))$. Then, for $n \geqq 1$, there exists a unique u in $H^{-\varepsilon}(\mathbb{R}; H)$, such that, for all v in $H^1(\mathbb{R}; W(\Omega_-))$, (4.37) holds.*

Proof. Define an operator $B_n(\omega)$ by

$$s(B_n(\tau)u, v) = \int L_n(k, \tau)\hat{u}(0, k) \cdot \bar{\hat{v}}(0, k)\, dk \quad \forall\, v \in W. \tag{4.38}$$

We solve (4.37) by Fourier transform in time:

$$i\omega \tilde{u}_n(\tau) + A\tilde{u}_n(\tau) + \mu \tilde{u}_n(\tau) + B_n(\tau)\tilde{u}_n(\tau) = u^0 + S\tilde{f}(\tau). \tag{4.39}$$

Here A and S are defined respectively at (3.59) and (3.61).

We estimate the solution of

$$i\omega\tilde{v}_n(\tau)+A\tilde{v}_n(\tau)+\mu\tilde{v}_n(\tau)+B_n(\tau)\tilde{v}_n(\tau)=v^0 \tag{4.40}$$

in terms of $\|v^0\|_H$ and $\|v^0\|_W$. If we multiply (4.40) scalarly by the conjugate of $\tilde{v}_n(t)$ and take the real part of the result, we obtain the estimate

$$\|\tilde{v}_n(\tau)\|_W \leqq C\|v^0\|_{W'}, \tag{4.41}$$

from the positivity of the operator B_n.

In order to obtain the $H^{-\varepsilon}$ estimate, we take the imaginary part of (4.40) scalarly multiplied by the conjugate of $\tilde{v}_n(\omega)$, and we obtain, for $n \geqq 1$,

$$\begin{aligned}
&\omega s(\tilde{v}_n(\tau), \bar{\tilde{v}}_n(\tau))+\left(a_1+\frac{2\tau\nu}{na_1}\right)\int N\hat{v}_n(0,k,\tau)\cdot\bar{\hat{v}}_n(0,k,\omega)\,dk \\
&\quad=\operatorname{Im}\left\{s(v^0,\bar{\tilde{v}}_n(\tau))-s(A\bar{\tilde{v}}_n(\tau),\tilde{v}_n(\tau))\right. \\
&\qquad-\int M\hat{v}_n(0,k,\tau)\cdot\bar{\hat{v}}_n(0,k,\tau)\,dk \\
&\qquad\left.+\int 2\nu\left[\lambda_n+\frac{i\omega'}{n}-i\frac{a_1}{na_2}\right]N\hat{v}_n(0,k,\tau)\cdot\bar{\hat{v}}_n(0,k,\tau)\,dk\right\}.
\end{aligned} \tag{4.42}$$

The most interesting term in the right-hand side of the above relation is the integral term which has $(\lambda_n+i\omega'/n)$ as a factor of the integrand; to estimate this term, we use Proposition 4.6. We will have a result if we are able to estimate $k\hat{v}_n(0,\cdot,\omega)$. This will depend on additional regularity on $\tilde{v}_n(\tau)$. Let D denote the differentiation with respect to x_2; if we apply D to (4.40), we can write, because $B_n(\tau)$ commutes obviously with D,

$$i\omega D\tilde{v}_n(\tau)+\mu\tilde{v}_n(\tau)+AD\tilde{v}_n(\tau)+B_n(\tau)D\tilde{v}_n(\tau)=Dv^0. \tag{4.43}$$

From (4.41), we have

$$\|D\tilde{v}_n(\tau)\|_W \leqq C\|Dv^0\|_{W'} \leqq C\|v^0\|_H.$$

Now, this inequality enables us to conclude that, for $\omega \neq 0$,

$$|\omega| s(\tilde{v}_n(\tau),\tilde{v}_n(\tau)) \leqq C(\|v^0\|_H)^2. \tag{4.44}$$

This relation concludes the proof of the existence. The uniqueness is as in the proof of Theorem 3.12. □

There is an analogous result for $n=0$; as it is easier, its proof is left to the reader.

PROPOSITION 4.9. *Let u^0 belong to W'; then, there exists a unique u in $H^{-\varepsilon}(\mathbb{R}; H)$ such that, for all v in $H^1(\mathbb{R}; W)$, (4.34) holds.*

These $H^{-\varepsilon}$ estimates are of course very bad, but we will obtain better after we make error estimates. Let us denote

$$e_n = u_n - u. \tag{4.45}$$

THEOREM 4.10. *Assume that u^0 and f satisfy the support conditions* (2.36) *and* (2.37); *then, for $n \geqq 1$, and for all positive p, the error e_n satisfies the estimate*

$$\|e_n\|_{H^p(\mathbb{R};\,W)} \leqq C'_{p,n}\nu^{2n}(\|u^0\|_H+\|f\|_{L^2(\mathbb{R};\,H)}). \tag{4.46}$$

For $n=0$, the exponent $2n$ has to be replaced by 1 in relation (4.46).

Proof. We subtract (3.63) from (4.39), and we obtain

$$i\omega\tilde{e}_n+A\tilde{e}_n+\mu\tilde{e}_n+B_n(\tau)\tilde{e}_n=(B-B_n)(\tau)\tilde{u}. \tag{4.47}$$

The point now is to estimate the norm of $(B-B_n)(\tau)\tilde{u}(\tau)$ in $W'(\Omega_-)$, so as to utilize (4.38). For $n \geqq 1$, we deduce from (4.8) and (4.32) that

$$s((B-B_n)(\tau)\tilde{u}(\tau), v) = \int \nu(\lambda_n - \lambda) N\hat{u}(0, \cdot, \tau) \cdot \hat{v}(0, \cdot, \tau)\, dk.$$

We already know from (4.10) that

$$|\lambda_n - \lambda| \leqq C^{n-1}\nu^{2n-2}|\lambda_1 - \lambda|(|k|^2+|\tau|^2)^{n-1}.$$

An easy algebraic computation shows that there exists a constant c' such that

$$|\lambda_1 - \lambda| \leqq c'\nu(|k|^2+|\tau|^2).$$

On the other hand, under the support conditions, $\mathcal{N}u|_\Sigma$ belongs to $H^\infty(\mathbb{R}^2)$. Thus, for all p, there exists a constant c_p such that

$$|N\hat{u}(0, k, \tau)| \leqq c_p(1+|k|^2+|\tau|^2)^{-p}.$$

These considerations imply the estimate

$$|s((B-B_n)(\tau)\tilde{u}(\tau), v)| \leqq C^{n-1}\nu^{2n}c_p c' \int (|k|^2+|\tau|^2)^n(1+|k|^2+|\tau|^2)^{-p}|\hat{v}(k)|\, dk.$$

If v is taken in W, then $\hat{v}$ is integrable on Σ, and there is a constant $C_{p,n}$ such that

$$|s((B-B_n)(\tau)\tilde{u}(\tau), v)| \leqq C_{p,n}\nu^{2n}\|v\|_W(1+|\tau|^2)^{p-n}.$$

Thus, we obtain

$$\|(B-B_n)(\tau)\tilde{u}(\tau)\|_{W'} \leqq C_{p,n}\nu^{2n}(1+|\tau|^2)^{p-n}.$$

From here, the conclusion of the theorem is immediate. The case $n=0$ is completely analogous. □

Remark 4.11. The constant $C_{p,n}$ which appears in (4.46) increases with n; moreover, the sequence λ_n converges to λ only for bounded values of λ and ω. If one wanted to prove that the limit of the sequence u_n is u, one would have to estimate an integral involving arbitrary powers of (ω'^2+k^2). Therefore, it is likely that a convergence theorem would need very strong conditions over the data.

COROLLARY 4.12. Under the assumptions of Theorem 4.10, u_n is an element of $L^\infty(\mathbb{R}^+, H(\Omega_-)) \cap L^2(\mathbb{R}^+; W(\Omega_-))$.

Proof. From Theorem 3.12, we know that u is an element of $L^\infty(\mathbb{R}^+, H(\Omega_-)) \cap L^2(\mathbb{R}^+; W(\Omega_-))$; as the error e_n belongs to the same space, the corollary holds.

In order to get rid of the support conditions, we prove some positivity results for $\mathcal{L}_n$.

LEMMA 4.13. The operator $\mathcal{L}_n$ is causal; if g is an element of $H^\infty(\mathbb{R}^2)$, which vanishes for t lesser than or equal to zero, then

$$\text{Re} \int_0^T \langle \mathcal{L}_n g, g\rangle_\Gamma\, dt \geqq 0. \tag{4.48}$$

Proof. The causality of $\mathcal{L}_n$ comes from the inductive construction of λ_n; it suffices to observe that λ_1 admits an extension to the half-plane $\text{Im}(\tau) < 0$, and that this extension has polynomial growth. By induction, the same holds for the λ_n. Details of this proof are left to the reader. The second part of the lemma relies on the decomposition

$$\mathcal{L}_n = \mathcal{R}_n + \frac{1}{n}\frac{d}{dt}, \tag{4.49}$$

where the symbol of $\mathcal{R}_n$ is bounded with respect to τ. This decomposition is an immediate consequence of Lemma 4.6. Arguing as in the proof of Lemma 3.7, and with the help of a technique used in [21], we define $\mathcal{L}_n^{\#}(k, \cdot)$ by

$$\mathcal{L}_n^{\#}(k, \cdot)u(t) = \mathcal{F}_{\omega \to t}\{L_n(k, \tau)\hat{u}(\tau)\}$$

and similarly $N^{\#}(k, \cdot)$, recalling the definition (3.50) of $\mathcal{N}$.

If we let

$$u_\varepsilon(t) = \begin{cases} u(t) \text{ if } 0 \leqq t \leqq T, \\ u(T)(T - t + \varepsilon)\varepsilon^{-1} \quad \text{if } T \leqq t \leqq T + \varepsilon, \\ 0 \text{ if } T + \varepsilon \leqq t, \end{cases}$$

then

$$\operatorname{Re} \int \mathcal{L}_n^{\#}(k, \cdot)u_\varepsilon(t) \cdot u_\varepsilon(t)\, dt = \operatorname{Re} \int L_n(k, \tau)\hat{u}_\varepsilon(\tau) \cdot \bar{\hat{u}}_\varepsilon(\tau)\, d\omega \geqq 0.$$

On the other hand, by causality,

$$\operatorname{Re} \int_0^T \mathcal{L}_n^{\#}(k, \cdot)u(t) \cdot u(t)\, dt = \operatorname{Re} \int \mathcal{L}_n^{\#}(k, \cdot)u_\varepsilon(t) \cdot u_\varepsilon(t)\, dt - \operatorname{Re} \int_T^{T+\varepsilon} \mathcal{L}_n^{\#}(k, \cdot)u_\varepsilon(t) \cdot u_\varepsilon(t)\, dt.$$

The first term is greater than or equal to zero; we estimate the second one as ε tends to zero; with the help of decomposition (4.49),

$$\lim_{\varepsilon \to 0} \operatorname{Re} \int_T^{T+\varepsilon} \mathcal{R}_n^{\#}(k, \cdot)u_\varepsilon(t) \cdot u_\varepsilon(t)\, dt = 0,$$

because $\mathcal{R}_n$ is bounded with respect to τ;

$$\lim_{\varepsilon \to 0} \operatorname{Re} \int_T^{T+\varepsilon} \mathcal{N}^{\#}(k, .) \frac{du_\varepsilon(t)}{dt} \cdot u_\varepsilon(t)\, dt = -\frac{1}{2}\mathcal{N}^{\#}(k, .)u(T) \cdot u(T),$$

as is shown by a straightforward computation. This shows the lemma. □

Now we can give a much more precise estimate under the assumptions of Theorem 4.10.

COROLLARY 4.14. *Under the assumptions of Theorem* 4.10, *the solution* u_n *with absorbing boundary conditions satisfies the estimate*

$$\frac{1}{2}s(u_n(t), u_n(t)) + \int_0^T \{\mu s(u_n(t), u_n(t)) + \tilde{\mathbf{a}}(u_n(t), u_n(t))\}\, dt$$

(4.50)

$$\leqq \frac{1}{2}s(u^0, u^0) + \int_0^T (f(t), u_n(t))\, dt.$$

Proof. This result is an immediate consequence of Lemma 4.13; one has only to substitute v by u_n in the variational equality (4.37). We can do this for smooth enough data; if u^0 and f are smooth enough, u_t belongs to $L^2(\mathbb{R}^+; H(\Omega_-))$. By error estimate (4.43), u_n belongs to the same space. Therefore, using the integration over $[0, T]$, the positivity and a density argument, one obtains the desired result. □

Finally, we obtain the most general existence and uniqueness theorem.

THEOREM 4.15. *For all* u^0 *in* $H(\Omega_-)$ *and all* f *in* $L^2(0, \infty; \mathbf{L}^2(\Omega_-))$, *and for all strictly positive* μ *there exists a unique* u *in* $L^\infty(\mathbb{R}^+; H(\Omega_-)) \cap L^2(\mathbb{R}^+; W(\Omega_-))$ *such that* u_t *belongs to* $L^2(\mathbb{R}; W'(\Omega_-))$ *and the variational equality* (4.37) *and the energy inequality* (4.50) *hold.*

Proof. It is enough to approximate any initial data by initial data satisfying the support condition. Then, the energy estimate gives the existence by standard procedure of extraction of subsequences. The uniqueness depends only on the positivity and still holds. □

4.3. Explicit formulations. We write problem (4.37) explicitly for $n=0$ and $n=1$. The variational form will be convenient for computations. We give the associated boundary conditions; for $n=2$, we use an auxiliary unknown.

Let us recall first that a product in Fourier space corresponds to a convolution in physical space. One of the important operators is the convolution by the inverse Fourier transform of $\mathrm{pf}(1/|k|)$; it is defined by

$$\mathcal{K}u(y)=\bar{\mathcal{F}}(u(k)pf(1/|k|)).$$

The kernel K of $\mathcal{K}$ is given by

$$K(x)=\frac{1}{\pi}(\gamma-\mathrm{Log}\,|x|), \tag{4.51}$$

where $\gamma\simeq 0.57721\ldots$ is the Euler constant. Then, the scalar product s admits the expression

$$\begin{cases} s(u,v)=\displaystyle\int_{\Omega_-} u(x_1,x_2)v(x_1,x_2)\,dx_1\,dx_2 \\ +\dfrac{1}{\pi}\left(\gamma\displaystyle\int_\Gamma u_1(0,x_2)v_1(0,x_2)\,dx_2-\barint_{\Gamma\times\Gamma}\mathrm{Log}\,|x_2-y_2|u_1(0,x_2)v_1(0,y_2)\,dx_2\,dy_2\right) \end{cases} \tag{4.52}$$

where the barred integral means that a principal value has to be taken.

Another important kernel is the kernel of the Hilbert transform, which is equal to $vp(1/\pi x)$. This distribution is not locally square integrable, but from the relation

$$\langle\mathcal{H}\varphi,\psi\rangle=-\frac{1}{2\pi}\int_{\mathbb{R}} i\sigma\hat{\varphi}\bar{\hat{\psi}}\,dk=\frac{1}{2\pi}\int_{\mathbb{R}}\frac{1}{|k|}\,\overline{\hat{\varphi}ik}\,\bar{\hat{\psi}}\,dk.$$

We deduce

$$\langle\mathcal{H}\varphi,\psi\rangle=\left\langle\mathcal{H}\varphi,\frac{\partial\psi}{\partial x_2}\right\rangle, \tag{4.53}$$

and the kernel K of $\mathcal{K}$ is locally integrable.

According to Definition 4.7, the symbol of $\mathcal{L}_0$ is the matrix L_0 given by

$$L_0(k,\tau,\nu)=M(k,\tau,0)+\frac{r_0}{2}N(k).$$

the value of r_0 is given by (4.6) and (4.9); $r_0=a_1$. Therefore,

$$L_0(k,\tau,\nu)=\begin{pmatrix} ia_2\sigma+(a_1/2) & -i\sigma a_1 \\ 0 & a_1/2 \end{pmatrix}. \tag{4.54}$$

Therefore,

$$(L_0\hat{u},\hat{v})=\frac{a_1}{2}(\hat{u},\hat{v})_\Gamma+(-i\sigma(a_1\hat{u}_2-a_2\hat{u}_1),\hat{v}_1)_\Gamma.$$

If the vector product is denoted by the symbol $\times$, this last expression can be rewritten as

$$(L_0\hat{u}, \hat{v}) = \frac{a_1}{2}(\hat{u}, \hat{v})_\Gamma + (-i\sigma\, a \times \hat{u}, \hat{v}_1)_\Gamma.$$

We obtain the following expression of $(\mathscr{L}_0 u, v)$:

$$(\mathscr{L}_0 u, v) = \frac{a_1}{2}(u, v)_\Gamma + (a \times \mathscr{H}u, v_1)_\Gamma,$$

or, with the help of (4.53),

$$(\mathscr{L}_0 u, v) = \frac{a_1}{2}(u, v)_\Gamma + \left(a \times \mathscr{H}u, \frac{\partial v_1}{\partial x_2}\right)_\Gamma. \tag{4.55}$$

Finally, the variational formulation for 0-artificial conditions is stated in the following proposition.

PROPOSITION 4.16. *For $n = 0$, the variational formulation* (4.37) *is equivalent to*

$$s(u, v) = (u, v)_{\Omega_-} + (\mathscr{H}u, v)_\Gamma; \tag{4.56}$$

$$s(u_t, v) + \mu s(u, v) + \tilde{\mathbf{a}}(u, v) + \frac{a_1}{2}(u, v)_\Gamma + \left(a \times \mathscr{H}u, \frac{\partial v_1}{\partial x_2}\right)_\Gamma = (f, v) \qquad \forall v \in W(\Omega_-). \tag{4.57}$$

Proof. The above dictionary of kernels proves that the following variational equality holds:

$$\int_{\mathbb{R}} \left[-s(u, v_t) + \mu s(u, v) + \tilde{\mathbf{a}}(u, v) + \frac{a_1}{2}(u, v)_\Sigma + \left(a \times \mathscr{H}u, \frac{\partial v_1}{\partial x_2}\right)_\Sigma \right] dt$$

$$= \int_{\mathbb{R}} (f, v)\, dt + s(u^0, v(0)).$$

for all test functions v in $L^2(\mathbb{R};\, W(\Omega_-))$ such that v_t belongs to $L^2(\mathbb{R};\, H(\Omega_-))$. Observe that the second integral along Γ makes sense because of the characterization of the trace on Γ. Using a test function of the form

$$v = W \otimes \varphi,$$

where v belongs to $W(\Omega_-)$ and φ is a smooth function, we obtain (4.57). □

In the case $n = 1$, L_1 is more complicated:

$$L_1(k, \tau, \nu) = M(k, \tau, \nu) + \frac{r_1}{2} N(k),$$

r_1 is given by

$$r_1 = a_1 + \frac{2\nu i}{a_1}(\tau + a_2 k).$$

Therefore,

$$L_1(k, \tau, \nu) = L_0(k, \tau, \nu) + \nu|k| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{\nu i}{a_1}(\tau + a_2 k) N(k). \tag{4.58}$$

Therefore, we have only to compute the extra terms which did not appear before:

$$\frac{1}{2\pi} \int_{\mathbb{R}} |k| \hat{u} \cdot \bar{\hat{v}}\, dk = \left(\mathscr{H} \frac{\partial u}{\partial x_2}, \frac{\partial v}{\partial x_2}\right)_\Gamma.$$

In the same fashion,

$$\frac{1}{2\pi}\int_{\mathbb{R}} ik\,N(k)\hat{u}\cdot\bar{\hat{v}}\,dk=\frac{1}{2\pi}\int_{\mathbb{R}} ik\hat{u}\cdot\bar{\hat{v}}\,dk+\frac{1}{2\pi}\int_{\mathbb{R}}|k|\bar{\hat{v}}\times\hat{u}\,dk$$
$$=\left(u,\frac{\partial v}{\partial x_2}\right)_\Gamma+\left(\mathcal{K}\frac{\partial v}{\partial x_2}\times\frac{\partial u}{\partial x_2},1\right)_\Gamma,$$
$$\frac{1}{2\pi}\int_{\mathbb{R}} N(k)\hat{u}\cdot\bar{\hat{v}}\,dk=\frac{1}{2\pi}\int_{\mathbb{R}}\hat{u}\cdot\bar{\hat{v}}\,dk+\frac{1}{2\pi}\int_{\mathbb{R}}(-i\sigma)\bar{\hat{v}}\times\hat{u}\,dk$$
$$=(u,v)_\Gamma+(\mathcal{K}v\times u,1)_\Gamma.$$

We can summarize this calculation:

$$(\mathcal{L}_1u,v)=(\mathcal{L}_0u,v)+\nu\left(\mathcal{K}\frac{\partial u}{\partial x_2},\frac{\partial v}{\partial x_2}\right)_\Gamma+\frac{\nu a_2}{a_1}\left(u,\frac{\partial v}{\partial x_2}\right)_\Gamma$$
$$+\left(\mathcal{K}\frac{\partial v}{\partial x_2}\times\frac{\partial u}{\partial x_2},1\right)_\Gamma+\frac{\nu}{a_1}(u_t+\mu u,v)_\Gamma+(\mathcal{K}v\times(u_t+\mu u),1)_\Gamma.$$

Finally, we have the following result:

PROPOSITION 4.17. *For $n=1$, the variational formulation* (4.37) *is equivalent to*

$$s(u_t,v)+\frac{\nu}{a_1}(u_t,v_t)_\Gamma+(\mathcal{K}v\times u_t,1)_\Gamma+\mu s(u,v)+\tilde{\mathbf{a}}(u,v)+\frac{a_1}{2}(u,v)_\Gamma$$
$$+\left(a\times\mathcal{K}u,\frac{\partial v_1}{\partial x_2}\right)_\Gamma+\nu\left(\mathcal{K}\frac{au}{\partial x_2},\frac{\partial v}{\partial x_2}\right)_\Gamma+\frac{\nu a_2}{a_1}\left(u,\frac{\partial v}{\partial x_2}\right)_\Gamma+\left(\mathcal{K}\frac{\partial v}{\partial x_2}\times\frac{\partial u}{\partial x_2},1\right)_\Gamma \tag{4.59}$$
$$=(f,v)\quad\forall\, v \text{ in } W(\Omega_-).$$

Proof. This is simply a consequence of the previous computation. □

In order to obtain the strong formulation of the boundary conditions, we observe that if the matrix E_m is defined by

$$E_m=\frac{a_1 I}{2}-K_1-L_m, \tag{4.60}$$

where K_1 is given by (3.44), then the associated operator $\mathscr{E}_m$ gives the boundary condition

$$\sigma_n=\mathscr{E}_m u|_\Sigma. \tag{4.61}$$

The matrix E_0 is given by

$$E_0=\begin{pmatrix} ia_2\sigma-(i\tau/|k|) & ia_1\sigma \\ 0 & 0\end{pmatrix}.$$

Therefore, the corresponding boundary condition is given by

$$\sigma_{11}=-\mathcal{K}\frac{\partial u_1}{\partial t}+\mathcal{H}u\times a, \tag{4.62}$$

$$\sigma_{12}=0. \tag{4.63}$$

The same kind of computation gives the result for $n=1$:

$$\sigma_{11}=-\mathcal{K}\frac{\partial u_1}{\partial t}+\mathcal{H}u\times a+\nu\mathcal{H}\frac{\partial u_1}{\partial x_2}-\frac{\nu}{a_1}\left(\frac{\partial}{\partial t}+a_2\frac{\partial}{\partial x_2}\right)(u_1+\mathcal{H}u_2), \tag{4.64}$$

$$\sigma_{21}=\nu\mathcal{H}\frac{\partial u_2}{\partial x_2}+\frac{\nu}{a_1}\left(\frac{\partial}{\partial t}+a_2\frac{\partial}{\partial x_2}\right)(-\mathcal{H}u_1+u_2). \tag{4.65}$$

In [13], the family λ_n defined by (4.4)–(4.8) had been introduced to design absorbing boundary conditions for the advection-diffusion equation $u_t + a\nabla u - \nu\Delta u = 0$. It turned out that with this special choice, the boundary conditions assumed a very special form, namely

$$\left(\frac{\partial}{\partial t} + a\nabla\right)^n u = 0.$$

The analysis is much more intricate here, and we will merely outline a possible strategy for the case $n = 2$. It relies on the introduction of an unknown auxiliary, defined on the boundary. This technique had been initiated in [23] and proved to be very useful [3]. Using the symbols and the expression of λ_2 and λ_1, we get a formulation of L_2 as

$$L_2 = L_1 + \nu(\lambda_1 - \lambda_2)N, \quad \text{or} \quad L_2 = L_1 - \nu^2 \frac{k^2 + \omega'^2}{1 + 2i\omega'\nu} N.$$

Let us introduce the auxiliary function φ defined on the boundary through its Fourier transform by the expression

$$-\nu \frac{k - i\omega'}{1 + 2i\omega'} N\hat{u} = \hat{\varphi}, \tag{4.66}$$

or equivalently, using the fact that $(Nu)_1 - \mathcal{H}(Nu)_2 = 0$, $\varphi_1 + \mathcal{H}\varphi_2 = 0$;

$$\left(1 + \frac{2\nu}{a_2}\left(\frac{\partial}{\partial t} + a_2 \frac{\partial}{\partial x_2}\right)\right)(\varphi_1 + \mathcal{H}\varphi_2) + 2\nu\mathcal{H}\left(i\frac{\partial}{\partial x_2} + \frac{1}{a}\left(\frac{\partial}{\partial t} + a_2\frac{\partial}{\partial x_2}\right)\right)(u_1 + \mathcal{H}u_2) = 0. \tag{4.67}$$

then $L_2\hat{u}$ is given by

$$L_2\hat{u} = L_1\hat{u} + \nu(k + i\omega')\hat{\varphi}. \tag{4.68}$$

We conclude that the boundary condition associated with the second approximation $\mathscr{L}_2$ reads

$$\begin{aligned}\sigma_{11} = &-\mathscr{K}\frac{\partial u_1}{\partial t} + \mathcal{H}u \times a + \nu\mathcal{H}\frac{\partial u_1}{\partial x_2} \\ &-\frac{\nu}{a_1}\left(\frac{\partial}{\partial t} + a_2\frac{\partial}{\partial x_2}\right)(u_1 + \mathcal{H}u_2 + \varphi_1) + i\nu\frac{\partial \varphi_1}{\partial x_2},\end{aligned} \tag{4.69}$$

$$\sigma_{21} = \nu\mathcal{H}\frac{\partial u_2}{\partial x_2} - \frac{\nu}{a_1}\left(\frac{\partial}{\partial t} + a_2\frac{\partial}{\partial x_2}\right)(u_2 - \mathcal{H}u_1 + \varphi_2) + i\nu\frac{\partial \varphi_2}{\partial x_2} = 0. \tag{4.70}$$

These formulae have to be supplemented with (4.67).

Remark 4.12. The first boundary condition (4.62), (4.63) is actually local and can be written in the form:

$$\frac{\partial u_1}{\partial t} + (a \cdot \nabla)u_1 = 0 \tag{4.71}$$

$$\frac{\partial u_2}{\partial x_1} = 0. \tag{4.72}$$

Appendix A. The standard spaces of Beppo-Levi functions are studied in [7]; the distribution spaces described by Definition 2.1 are of the very same nature. The common property of the spaces $BL(H^s(\Omega))$ is that their local properties are the same as the local properties of $H^{s+1}(\Omega)$, but their properties in the large are quite different. In particular,

LEMMA A.1. *For any n, for any unbounded open set* Ω *of* $\mathbb{R}^n$, *and for any real s, there exists an unbounded function in* $BL(H^s(\Omega))$.

Proof. If is enough to exhibit examples; outside of a compact subset, we require u to be equal to

$$r^{1/4} \text{ if } n=1;$$

$$\text{Log Log } r \text{ if } n=2;$$

$$r^{-1/4} \text{ if } n \geqq 3.$$

Checking that these functions give the answer is an exercise left to the reader.

Though the elements of $BL(H^s(\Omega))$ may be unbounded, they are nonetheless elements of $\mathcal{S}'$. This is a consequence of the fact that $vp(1/ik)$ in dimension 1, and proper generalizations of this in dimension greater than or equal to 2 admit a Fourier transform, which is known explicitly. Moreover, the growth of the elements of $BL(H^s(\Omega))$ is polynomial at most and can be estimated precisely.

Finally, the Beppo-Levi spaces are natural spaces on an unbounded domain, where the only estimate is an energy estimate that involves only the gradient.

Appendix B. In this appendix, we prove a result on the behavior at infinity of harmonic functions. This result is probably part of the folklore of the subject, but we know no source where to find it in simple form.

LEMMA B.1. *Let p be a function on* $\mathbb{R}^N$ *such that* Δp (*computed in the sense of distributions*) *has compact support. If p decreases fast at infinity to zero, then p is constant outside of a compact set of* $\mathbb{R}^N$.

Proof. Let ρ be a C^∞ test function that is radial and has support in the ball of center 0 and radius 1. Denote $\rho_\varepsilon(x)=\varepsilon^{-N}\rho(x/\varepsilon)$. If p is harmonic for all x in $E_{R-2}=\{x/|x|\geqq R-2\}$, then it has the mean property in this region, and, whenever $|x|\geqq R-2+\varepsilon$, we have

$$u(x)=u*\rho_\varepsilon(x).$$

Therefore, u has derivatives of all orders in int (E_{R-2}), and all of its derivatives decrease rapidly at infinity. Let φ be an infinitely differentiable function on $\mathbb{R}^N$ such that

$$\varphi=0 \text{ for } |x|\leqq R-1$$

$$\varphi=1 \text{ for } |x|\geqq R.$$

Then $q=p\varphi$ is in $\mathcal{S}(\mathbb{R}^N)$, and

$$\psi=\Delta q \text{ has support in the ball of center 0 and radius } R.$$

We perform a Fourier transform on the equation $\Delta q=\psi$, and we obtain

$$|\xi|^2\hat{q}(\xi)=\hat{\psi}(\xi). \tag{B.1}$$

Thanks to the Paley-Wiener theorem, $\hat{\psi}(\xi)$ can be extended to all of $\mathbb{C}^N$ and is an entire function of ξ; moreover, for all n, there is a constant C_n such that

$$|\hat{\psi}(\xi)|\leqq C_n(1+|\xi|)^n e^{R|\mathrm{Im}(\xi)|}.$$

From (B.1), we can see that $\hat{q}$ can be extended to all $\mathbb{C}^N$ as a meromorphic function of ξ, with possibly a pole at zero. But, since q is in $\mathcal{S}(\mathbb{R}^N)$, $\hat{q}$ is in $\mathcal{S}(\mathbb{R}^N)$, too, and there is no pole of $\hat{q}$ at zero. Thus $\hat{q}$ satisfies the estimate

$$|\hat{q}(\xi)|\leqq C_0 \quad \text{for } |\xi|\leqq 1,$$

$$|\hat{q}(\xi)|\leqq C_n(1+|\xi|)^n e^{R|\mathrm{Im}(\xi)|}, \quad \text{for } |\xi|\geqq 1.$$

Thus q has compact support in the ball $|x| \leqq R$, and therefore p vanishes for $|x| \geqq R$. □

If p is known in (2.29)-(2.31), then u is the solution of an advection diffusion with right-hand side $f - \nabla p$; therefore, if p does not decrease rapidly to zero at infinity, and if f has compact support, for instance, u cannot generally decrease rapidly to zero at infinity. Therefore, a smooth solution u is in a space of function with polynomial estimates at infinity, and the dual of this space is a strict subspace of $\mathcal{S}'$.

REFERENCES

[1] G. ALLAIN, *Thèse de troisième cycle*, Université Pierre et Marie Curie, Paris, 1986.

[2] K. J. BAI, *A variational method in potential flows with a free surface*, College of Engineering, University of California, Berkeley, CA, Report 72-2, 1972.

[3] A. BAMBERGER, B. ENGQUIST, L. HALPERN, AND P. JOLY, *Higher-order paraxial wave equation approximation in heterogeneous media*, SIAM J. Appl. Math., 48 (1988), pp. 129-154.

[4] A. BAYLISS AND E. TURKEL, *Radiation boundary conditions for wave-like equations*, Comm. Pure Appl. Math., 33 (1980), pp. 707-725.

[5] ———, *Far field boundary conditions for compressible flow*, J. Comput. Phys., 48 (1982), pp. 182-199.

[6] J. L. LIONS AND R. DAUTRAY, *Analyse mathématique et calcul numérique pour les sciences et les techniques*. Tome 2, Masson, Paris, 1985.

[7] J. DENY AND J. L. LIONS, *Les espaces du type Beppo-Levi*, Ann. Inst. Fourier (Prenoble), 5 (1955), pp. 305-370.

[8] B. ENGQUIST AND A. MAJDA, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 139 (1977), pp. 629-651.

[9] ———, *Radiation boundary conditions for acoustic and elastic wave calculations*, Comm. Pure Appl. Math., 32 (1979), pp. 313-357.

[10] L. FERM AND B. GUSTAFSSON, *A down-stream boundary procedure for the Euler equations*, Comput. and Fluids, 10 (1982), pp. 261-276.

[11] T. M. HAGSTRÖM, *Boundary conditions at outflow for a problem with transport and diffusion*, J. Comput. Phys., 69 (1987), pp. 69-80.

[12] T. M. HAGSTRÖM AND H. B. KELLER, *The numerical calculation of traveling wave solutions of nonlinear parabolic solutions*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 978-988.

[13] L. HALPERN, *Artificial boundary conditions for the linear advection-diffusion equations*, Math. Comp., 46 (1986), pp. 425-438.

[14] ———, Approximations paraxiales et conditions aux limites absorbantes, Thèse d'Etat, Université Pierre et Marie Curie, Paris, 1986.

[15] L. HALPERN AND M. SCHATZMAN, *Conditions aux imites artificielles pour les équations de Navier-Stokes incompressibles*, C.R. Acad. Sci. Paris Sér. I, 304 (1987), pp. 83-86.

[16] G. W. HEDSTRÖM, *Non-reflecting boundary conditions for nonlinear hyperbolic systems*, J. Comput., Phys., 30 (1979), pp. 222-237.

[17] L. HÖRMANDER, *Pseudo-differential operators and hypoelliptic equations*, Proc. Sympos. Pure Math., 10 (1966), pp. 138-183.

[18] C. JOHNSON AND J. C. NEDELEC, *On the coupling of boundary integral and finite element methods*, Math. Comp., 35 (1980), pp. 1063-1079.

[19] P. JOLY, *Pseudo-transparent boundary conditions for the diffusion equation*, preprint.

[20] H. KUMANO-GO, *Algebras of pseudo-differential operators*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 17 (1970), pp. 31-50.

[21] G. LEBEAU AND M. SCHATZMAN, *A wave problem in a half-space with a one-sided constraint on the boundary*, J. Differential Equations, 53 (1984), pp. 309-361.

[22] M. LENOIR AND A. TOUNSI, *The localized finite element method and its application to the* 2-D *seakeeping problem*, SIAM J. Numer. Anal., 25 (1988), pp. 729-752.

[23] E. L. LINDMANN, *Free space boundary conditions for the time dependent wave equation*, J. Comput. Phys., 21 (1976), pp. 251-269.

[24] J. L. LIONS AND E. MAGENES, *Problèmes aux limites nonhomogènes et applications* (Tome 1), Dunod, Paris, 1968.

[25] D. M. RUDY AND J. C. STRIKWERDA, *A nonreflecting outflow boundary condition for subsonic Navier-Stokes calculations*, J. Comput. Phys., 36 (1980), pp. 55-70.

[26] A. SEQUEIRA, *The coupling of boundary integral and finite element methods for the bidimensional steady Stokes problem*, Math. Methods Appl. Sci., 5 (1983), pp. 356-376.

[27] E. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton, 1977.

[28] R. TEMAM, *Navier-Stokes Equations, Theory and Numerical Analysis*, North-Holland, Amsterdam, New York, 1977.

[29] K. W. THOMPSON, *Time dependent boundary conditions for hyperbolic systems*, J. Comput. Phys., 68 (1987), pp. 1-24.

[30] A. UNTERBERGER AND J. BOKOBZA, *Sur une généralisation des opérateurs de Calderon-Zygmund et des espaces H^s*, C.R. Acad. Sci. Paris, 260 (1965), pp. 3265-3267.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 354-366, March 1989

004

BLOW-UP ESTIMATES FOR A NONLINEAR HYPERBOLIC HEAT EQUATION*

HAMID BELLOUT† AND AVNER FRIEDMAN‡

Abstract. Consider the Cauchy problem for

$$\varepsilon u_{tt} + u_t - u_{xx} = F(u);$$

u represents the temperature when the standard Fourier law $q = u_x$ (q flux) is relaxed and $F(u)$ is a nonlinear source of energy. It is established that the solution exists for $0 < t < \phi_\varepsilon(x)$, and it blows up as $t \to \phi_\varepsilon(x)$. Further, $\phi_\varepsilon(x) \to T_0$ as $\varepsilon \to 0$ where T_0 is the blow-up time for $u_t - u_{xx} = F(u)$.

Key words. blow-up of solutions, blow-up time, hyperbolic equations

AMS(MOS) subject classifications. 35L05, 35L67, 35L70

1. Introduction. Recently there has been increasing interest in the blow-up of solutions of nonlinear heat equations, such as

$$u_t - u_{xx} = F(u), \tag{1.1}$$

and nonlinear wave equations, such as

$$u_{tt} - u_{xx} = F(u); \tag{1.2}$$

typically $F(u) \sim Au^p$ ($p > 1$) or $F(u) \sim e^u$ as $u \to \infty$; see [10], [11], [13], [16], [17] and the references given there regarding (1.1), and [4], [5], [14] regarding (1.2).

Equation (1.1) models the heat equation when the flux q is given by the Fourier law $q = -u_x$ and the conservation of energy equation is

$$u_t + q_x = G \quad (G \text{ a source of energy}). \tag{1.3}$$

Fourier's law implies infinite velocity of heat propagation, and there have been a number of modified laws that rule out this feature. One common version is [1], [2], [3], [6], [7], [18], and [19]

$$q(x, t+\varepsilon) = -u_x(x, t) \quad (\varepsilon > 0)$$

or its approximation

$$q(x, t) + \varepsilon q_t(x, t) = -u_x(x, t). \tag{1.4}$$

The conservation of energy equation (1.3) then needs to be modified also (see [8]), but for ε small it is approximately the same as before. From (1.3), (1.4) we deduce

$$\varepsilon u_{tt} + u_t - u_{xx} = F \tag{1.5}$$

where $F = G + \varepsilon G_t$ ($G_t = \partial G(u(x, t))/\partial t$); for ε small, $F \approx G$.

* Received by the editors October 26, 1987; accepted for publication (in revised form) June 22, 1988. This research was partially supported by National Science Foundations grant DMS-8612880.

† Northern Illinois University, Department of Mathematics, De Kalb, Illinois 60115.

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

In this paper we study the Cauchy problem for (1.5) with $F = F(u)$. If

$$\begin{aligned} u(x,0) &= f(x), \\ u_t(x,0) &= g(x) \end{aligned} \tag{1.6}$$

then, since it is natural to assume that, initially, the temperature satisfies the heat equation, we are led to the assumption that

$$g = f_{xx} + F(f). \tag{1.7}$$

In § 2 we represent (1.5), (1.6) in two equivalent but useful forms. In § 3 we establish the existence of unique solution u_ε of (1.5), (1.6) for $0 < t < \phi_\varepsilon(x)$, which blows up to $+\infty$ as $t \to \phi_\varepsilon(x)$. The method of Caffarelli and Friedman [4], [5] shows that if $f(u) \sim u^p$ as $u \to \infty$ then $\phi_\varepsilon(x)$ is continuously differentiable. Additional estimates on u_ε (independent of ε) are established in § 4.

Next, in § 5 we prove that

$$\liminf_{\varepsilon \to 0} \phi_\varepsilon(x) \geqq T_0 \tag{1.8}$$

where T_0 is the blow-up time for (1.1) with $u(x,0) = f(x)$. Finally, in § 6 we prove that

$$\limsup_{\varepsilon \to 0} \phi_\varepsilon(x) \leqq T_0. \tag{1.9}$$

Some generalizations and extensions are given in § 7.

Results of the type (1.8), (1.9) have been established for other types of equations in [12], [14], [15].

Assumptions. Throughout this paper we assume that $F \in C^2$,

$$\begin{aligned} &F(x) \geqq 0, \qquad F''(s) \geqq 0 \quad \text{if } s \geqq 0, \\ &F'(s) > 0 \quad \text{if } s > 0, \\ &\int_1^\infty \frac{ds}{F(s)} < \infty; \end{aligned} \tag{1.10}$$

$$f \geqq 0, \qquad f + \varepsilon g \geqq 0 \quad \text{(for all small } \varepsilon > 0), \tag{1.11}$$

$$f, g \quad \text{belong to } C^3(\mathbb{R}), \tag{1.12}$$

$$\begin{aligned} &f(x) + |g(x)| \leqq \frac{C}{(1+|x|)^\alpha}, \\ &\sum_{i=1}^{3} [|f^{(i)}(x)| + |g^{(i)}(x)|] \leqq C \quad \text{for more constants } C > 0,\ \alpha > 0 \end{aligned} \tag{1.13}$$

and (1.7) holds. In § 6 we will also need the assumption

$$g \geqq 0. \tag{1.14}$$

The results of this paper extend also to the case where (1.7) is not valid and to space dimension $\leqq 3$ (under some additional assumptions on F); see § 7.

2. Equivalent formulation for the Cauchy problem. We denote by u_ε the solution (if existing) of

$$\begin{aligned} &\mathcal{L}_\varepsilon u \equiv \varepsilon u_{tt} + u_t - u_{xx} = F(u), \\ &u(x,0) = f(x), \\ &u_t(x,0) = g(x). \end{aligned} \tag{P_ε^1}$$

Setting

$$v(x,t)=u(x,t)\,e^{t/(2\varepsilon)} \tag{2.1}$$

we get the equivalent system

$$\varepsilon v_{tt}-v_{xx}=F(v\,e^{-t/(2\varepsilon)})\,e^{t/(2\varepsilon)}+\frac{1}{4\varepsilon}\,v,$$

$$(P_\varepsilon^2)\qquad v(x,0)=f(x),$$

$$v_t(x,0)=\frac{1}{2\varepsilon}f(x)+g(x).$$

Setting

$$w(x,\tau)=v(x,t)\quad\text{where }\tau=\frac{t}{\sqrt{\varepsilon}}, \tag{2.2}$$

we get

$$w_{\tau\tau}-w_{xx}=F(w\,e^{-\tau/(2\sqrt{\varepsilon})})\,e^{\tau/(2\sqrt{\varepsilon})}+\frac{1}{4\varepsilon}\,w,$$

$$(P_\varepsilon^3)\qquad w(x,0)=f(x),$$

$$w_\tau(x,0)=\frac{1}{2\sqrt{\varepsilon}}f(x)+\sqrt{\varepsilon}\,g(x).$$

The concept of a solution of (P_ε^i) is always understood in the classical sense. Using the representation formula

$$\begin{aligned}z(x,t)=&\frac{1}{2}[z(x+t,0)+z(x-t,0)]+\frac{1}{2}\int_{x-t}^{x+t}z_t(\xi,0)\,d\xi\\&+\frac{1}{2}\int_0^t ds\int_{x-t+s}^{x+t-s}(z_{tt}-z_{xx})(y,s)\,dy\end{aligned} \tag{2.3}$$

for $z=w$ we then obtain for u the representation:

$$\begin{aligned}u(x_0,t_0)=&\frac{e^{-t_0/(2\varepsilon)}}{2\sqrt{\varepsilon}}\iint_{K_-^\varepsilon(x_0,t_0)}e^{t/(2\varepsilon)}\mathscr{L}_\varepsilon(u)\,dx\,dt+\frac{e^{-t_0/(2\varepsilon)}}{8\varepsilon^{3/2}}\iint_{K_-^\varepsilon(x_0,t_0)}e^{t/(2\varepsilon)}u\,dx\,dt\\&+\frac{1}{2}\,e^{-t_0/(2\varepsilon)}\left[f\left(x_0+\frac{t_0}{\sqrt{\varepsilon}}\right)+f\left(x_0-\frac{t_0}{\sqrt{\varepsilon}}\right)\right]\\&+\frac{1}{2}\,e^{-t_0/(2\varepsilon)}\int_{x_0-t_0/\sqrt{\varepsilon}}^{x_0+t_0/\sqrt{\varepsilon}}\left[\frac{1}{2\sqrt{\varepsilon}}f(x)+\sqrt{\varepsilon}\,g(x)\right]dx,\end{aligned} \tag{2.4}$$

where $x_0\in\mathbb{R}$, $t_0>0$ and

$$K_-^\varepsilon(x_0,t_0)=\left\{(x,t)\in\mathbb{R}\times(0,\infty),\,|x-x_0|<\frac{t_0-t}{\sqrt{\varepsilon}}\right\}.$$

THEOREM 2.1. *Let* (1.7) *and* (1.10)-(1.13) *hold. Then there exist functions* $\phi_\varepsilon(x)$, $u_\varepsilon(x, t)$ *satisfying*

$$0 < c \leqq \phi_\varepsilon(x) \leqq \infty \quad \text{for some } c > 0 \quad (c \text{ independent of } \varepsilon); \tag{2.5}$$

$$\text{if } \phi_\varepsilon(x) \not\equiv \infty \text{ then } \phi_\varepsilon(x) < \infty \text{ for all } x \in \mathbb{R} \quad \text{and} \quad \frac{|\phi_\varepsilon(x) - \phi_\varepsilon(x')|}{|x - x'|} \leqq \sqrt{\varepsilon} \ \forall x, x'; \tag{2.6}$$

u_ε *is a solution of* (P_ε^1) *in the region*

$$\Omega_\varepsilon = \{(x, t) \in \mathbb{R} \times [0, \infty);\ y < \phi_\varepsilon(x)\}$$

and

$$u_\varepsilon \geqq 0, \tag{2.7}$$

$$u_\varepsilon(x, t) \to \infty \quad \text{if } t \to \phi_\varepsilon(x); \tag{2.8}$$

The pair $(\phi_\varepsilon, u_\varepsilon)$ *is uniquely determined.*

3. *Proof of Theorem* 2.1. Let

$$F_n(u) = \begin{cases} F(\min(u, n)) & \text{if } u > 0 \quad (n = 1, 2, \cdots) \\ F(0) & \text{if } u \leqq 0 \end{cases}$$

and denote by u_n the solution of (P_ε^1) corresponding to F_n. The corresponding $w = w_n$ satisfy:

$$w^n_{\tau\tau} - w^n_{xx} = F_n(w^n e^{-\tau/(2\sqrt{\varepsilon})})\, e^{\tau/(2\sqrt{\varepsilon})} + \frac{1}{4\varepsilon} w^n,$$

$$w^n(x, 0) = f(x), \tag{3.1}$$

$$w^n_\tau(x, 0) = \frac{1}{2\sqrt{\varepsilon}} f(x) + \sqrt{\varepsilon} g(x).$$

Using (1.11) in the representation (2.3) for w^n, we can deduce by a continuity argument that $w^n(x, t) \geqq 0$ for $x \in \mathbb{R}$ and all $t > 0$.

Next we apply the arguments in [4] to w^n to deduce that, for a subsequence $w_{n'}$, $w \equiv \lim w_{n'}$ exists if $0 \leqq \tau < \tilde{\phi}_\varepsilon(x)$ and $w \equiv \infty$ if $\tau > \tilde{\phi}_\varepsilon(x)$, and, if $\tilde{\phi}_\varepsilon \not\equiv +\infty$, $\tilde{\phi}_\varepsilon$ is Lipschitz continuous with coefficient 1. The last fact is based on the inequalities

$$w_\tau \geqq \pm w_x - c_0 \quad (c_0 \text{ constant})$$

whose proof is as in [4]. Next, instead of (1.22) in [4] we have

$$\frac{w_{n,\tau}}{F(w_n + C_0)} \leqq C \quad \text{for some } C_0 > 0,\ C > 0$$

and this implies (using the last condition in (1.10)) that

$$w(x, \tau) \to \infty \quad \text{it } \tau \to \tilde{\phi}_\varepsilon(x).$$

In view of (2.1), (2.2) we conclude that the corresponding subsequence $u_{n'}$ converge to u_ε which is finite if $t < \phi_\varepsilon(x)$ and $+\infty$ if $t > \phi_\varepsilon(x)$, where $\phi_\varepsilon(x) = \sqrt{\varepsilon}\tilde{\phi}_\varepsilon(x)$, and $u_\varepsilon(x, t) \to \infty$ if $t \to \phi_\varepsilon(x)$.

We have thus established (2.6)-(2.8). To prove the first inequality in (2.5) it suffices to establish the following lemma.

LEMMA 3.1. *There exist positive constants* M, T *independent of* n, ε *such that*

$$\sup_{\mathbb{R}^n \times (0, T)} u^n(x, t) \leqq M. \tag{3.2}$$

Proof. We compare u^n with the solution $\gamma(t)=\gamma_n(t)$ of

$$\begin{aligned} &\varepsilon\gamma''+\gamma'=F_n(\gamma), \\ &\gamma(0)=a, \\ &\gamma'(0)=F_n(a) \end{aligned} \tag{3.3}$$

where a is sufficiently large positive constant. The functions $z^n=u^n-\gamma_n$ satisfy:

$$\mathscr{L}_\varepsilon(z^n)=F_n(u^n)-F_n(\gamma_n)=c(x,t)z^n, \qquad c\geqq 0,$$

$$z^n(x,0)<0, \qquad z_t^n(x,0)\leqq 0.$$

Representing z^n by the integral formula (2.3) we can establish by continuity in t that $z^n(x,t)<0$ for all x, t. Thus in order to complete the proof of Lemma 3.1 it remains to prove:

$$\gamma_n(T)\leqq 2a \quad \text{for some } T \text{ independent of } n, \varepsilon. \tag{3.4}$$

To prove (3.4) we rewrite the differential equation for $\gamma=\gamma_n$ in the form

$$\varepsilon(e^{t/\varepsilon}\gamma')'=F_n(\gamma)\,e^{t/\varepsilon}\geqq 0.$$

Since $\gamma'(0)>0$, we deduce that $\gamma'(t)>0$ as long as $\gamma(t)$ is positive. Differentiating the equation in (3.3) once in t, we also have

$$\varepsilon(e^{t/\varepsilon}\gamma'')'=F_n'(\gamma)\gamma'\,e^{t/\varepsilon}\geqq 0$$

and, since $\gamma''(0)=0$, we deduce that $\gamma''(t)>0$ as long as $\gamma(t)>0$. It follows that $\gamma''(t)>0$ for all t. Therefore

$$\gamma'=F_n(\gamma)-\varepsilon\gamma''\leqq F(\gamma).$$

Defining T_n by $\gamma(T_n)=2a$, we conclude that

$$\tilde{T}\equiv\int_a^{2a}\frac{ds}{F_n(s)}=\int_0^{T_n}\frac{\gamma'(t)}{F(\gamma(t))}\leqq T_n.$$

Since $T_n\geqq\tilde{T}>0$ and $\tilde{T}$ is independent of n if n is large enough, the assertion (3.4) follows.

We have completed the proof of existence of a solution $(u_\varepsilon, \phi_\varepsilon)$. Uniqueness now follows by an easy argument; see [4] or [5] for details. (Note that for uniqueness we need not use the fact that c in (2.5) is independent of ε.)

As in [4] we can establish that u_ε is in $C^{2,1}$ in Ω_ε.

The following fact will be used in § 6.

THEOREM 3.2. *If in addition to the assumptions* (1.7), (1.10)-(1.13) *we also assume that* (1.14) *holds, then*

$$\frac{\partial u_\varepsilon}{\partial t}\geqq 0 \quad \text{in } \Omega_\varepsilon. \tag{3.5}$$

Proof. Set

$$z(x,\tau)=u_t^n(x,t)\,e^{t/(2\varepsilon)}, \qquad \tau=\frac{t}{\sqrt{\varepsilon}}. \tag{3.6}$$

Then

$$z_{\tau\tau} - z_{xx} = F_n'(u^n)z + \frac{1}{4\varepsilon} z,$$

$$z(x, 0) = g(x), \tag{3.7}$$

$$z_\tau(x, 0) = \frac{1}{2\sqrt{\varepsilon}} g(x).$$

Consider first the case where $g(x) \geqq \delta > 0$. Then representing z by an integral formula (2.3) and proceeding by continuity on τ, we can establish that $z(x, \tau) > 0$ for all x, τ. Applying this to (3.7) with $g \geqq 0$ replaced by $g + \delta$, and letting $\delta \to 0$, yields $z \geqq 0$ where z is given by (3.6), and (3.5) is then proved by taking $n \to \infty$. □

4. Additional estimates on u_ε.

LEMMA 4.1. *Assume that for some positive constants M, T the solution of (P_1^ε) (established in Theorem 1.1) satisfies*:

$$u_\varepsilon(x, t) \leqq M \quad \text{in } \mathbb{R} \times [0, T], \quad \text{for all small } \varepsilon. \tag{4.1}$$

Then there exists a positive constant C_1 independent of ε such that

$$u_\varepsilon(x, t) + |u_{\varepsilon,t}(x, t)| \leqq \frac{C_1}{(1+|x|)^\alpha} \quad \text{in } \mathbb{R} \times [0, T], \tag{4.2}$$

$$|u_{\varepsilon,x}| + |u_{\varepsilon,t}| + |u_{\varepsilon,xx}| + |u_{\varepsilon,tx}| + |u_{\varepsilon,xxx}| + |u_{\varepsilon,txx}| \leqq C_1 \quad \text{in } \mathbb{R} \times [0, T]. \tag{4.3}$$

Proof. Consider the function

$$W_\alpha(x, t) = \frac{e^{At}}{(1+x^2)^{\alpha/2}}, \quad A > 0.$$

For any A (no matter how large), if ε is small enough then

$$\mathscr{L}_\varepsilon W_\alpha \geqq \frac{A}{2} W_\alpha.$$

On the other hand

$$\mathscr{L}_\varepsilon u_\varepsilon = \tilde{F}(u_\varepsilon) u_\varepsilon \quad \left(\tilde{F}(v) = \frac{F(v)}{v}\right)$$

and therefore the function $z = W - u_\varepsilon$ satisfies

$$\mathscr{L}_\varepsilon z \geqq \frac{A}{2} W_\alpha - \tilde{F}(u_\varepsilon) u_\varepsilon \geqq \tilde{F}(u_\varepsilon) z \quad \text{if } t < T$$

provided we choose A such that $A > 2\tilde{F}(m)$. Noting that

$$z(x, 0) > 0, \qquad z_t(x, 0) > 0$$

if A is large, we can represent z by the integral representation (2.3) and then deduce by continuity on t, that $z \geqq 0$ if $t < T$. Thus

$$u_\varepsilon \leqq W_\alpha = \frac{e^{AT}}{(1+x^2)^{\alpha/2}}.$$

Similarly, from the equation

$$\mathscr{L}_\varepsilon(u_{\varepsilon,t}) = F'(u_\varepsilon) u_{\varepsilon,t}$$

and the fact that $|F'(u_\varepsilon)| \leqq F'(M)$ we can proceed as before to estimate $u_{\varepsilon,t}$ from above by the same function W (with a different constant A). The function $-u_{\varepsilon,t}$ is estimated similarly. Thus (4.2) is proved.

The function $u_{\varepsilon,x}$ is estimated similarly using the comparison function W_α with $\alpha = 0$.

Next we differentiate $\mathscr{L}_\varepsilon u_\varepsilon = F$ once in t and once in x and obtain

$$\mathscr{L}_\varepsilon u_{\varepsilon,tx} = F'(u_\varepsilon)u_{\varepsilon,tx} + F''(u_\varepsilon)u_{\varepsilon,t}u_{\varepsilon,x}.$$

Noting that

$$|F''(u_\varepsilon)u_{\varepsilon,t}u_{\varepsilon,x}| \leqq C,$$

we can proceed as before to compare $u_{\varepsilon,tx}$ with $W_0 \equiv e^{At}$ provided A is sufficiently large. We thus obtain the estimate

$$|u_{\varepsilon,tx}| \leqq C_1; \quad C_1 \text{ constant.}$$

Similarly we establish the estimate

$$|u_{\varepsilon,xx}| \leqq C_1.$$

Differentiating $\mathscr{L}(u_\varepsilon) = F$ three times and using the estimates derived so far, we can again compare $u_{\varepsilon,xxx}$ and $u_{\varepsilon,txx}$ with W_0 and thus complete the proof of (4.3). □

Remark 4.1. Lemma 4.1 implies that any sequence $\varepsilon \to 0$ has a subsequence such that

$$\begin{aligned} &u_\varepsilon \to u, \quad u_{\varepsilon,x} \to u_x, \quad u_{\varepsilon,xx} \to u_{xx} \\ &\text{uniformly in compact subsets of } \mathbb{R} \times [0, T]. \end{aligned} \tag{4.4}$$

However, we cannot establish the boundedness of $u_{\varepsilon,tt}$ by the method of Lemma 4.1 (since $u_{\varepsilon,ttt}(x, 0)$ is unbounded as $\varepsilon \to 0$), and thus we cannot assert that

$$u_{\varepsilon,t} \to u_t \quad \text{uniformly in compact subsets of } \mathbb{R} \times [0, T]. \tag{4.5}$$

LEMMA 4.2. *Under the assumption of Lemma* 4.1

$$u_\varepsilon(x, t) \to u(x, t), \qquad u_{\varepsilon,t}(x, t) \to u_t(x, t) \tag{4.6}$$

as $\varepsilon \to 0$, *uniformly in compact subsets of* $\mathbb{R} \times (0, T]$, *where u is a solution of* (1.1).

Proof. Multiplying $\mathscr{L}_\varepsilon(u_\varepsilon) = F(u_\varepsilon)$ by $e^{t/\varepsilon}$ and integrating in t we find that

$$u_{\varepsilon,t}(x, t_0) = \frac{1}{\varepsilon} e^{-t/\varepsilon} g(x) + \frac{1}{\varepsilon} e^{-t/\varepsilon} \int_0^t [u_{\varepsilon,xx}(x, s) + F(u_\varepsilon(x, s))]\, e^{s/\varepsilon}\, ds.$$

Set

$$H_\varepsilon = u_{\varepsilon,xx} + F(u_\varepsilon)$$

and write

$$\begin{aligned} u_{\varepsilon,t} &= \frac{1}{\varepsilon} e^{-t/\varepsilon} g(x) + \frac{1}{\varepsilon} \int_0^t [H_\varepsilon(x, s) - H(x, t)]\, e^{s/\varepsilon}\, ds + H_\varepsilon(x, t)[1 - e^{t/\varepsilon}] \\ &\equiv J_1^\varepsilon + J_2^\varepsilon + J_3^\varepsilon. \end{aligned} \tag{4.7}$$

Then

$$|J_1^\varepsilon| \leqq \frac{c}{\varepsilon} e^{-t/\varepsilon} \to 0$$

uniformly in $t \in [\delta, T]$.

Next, by Lemma 4.1, $|H_{\varepsilon,t}| \leqq C_0$ where C_0 is independent of ε. Hence

$$|J_2^\varepsilon| \leqq \frac{C_0}{\varepsilon} \int_0^t (t-s)\, e^{(s-t)/\varepsilon}\, ds$$

$$= C_0 \varepsilon \left[-\frac{t}{\varepsilon} e^{-t/\varepsilon} - e^{-t/\varepsilon} + 1 \right] \leqq C_0 \varepsilon.$$

Finally, by Remark 4.1, any sequence $\varepsilon \to 0$ has a subsequence such that (4.4) holds; therefore

$$J_3^\varepsilon \to u_{xx} + F(u).$$

Thus, by (4.7),

$$u_{\varepsilon,t} \to u_{xx} + F(u)$$

uniformly in compact subsets of $\mathbb{R} \times (0, T]$; the right-hand side must coincide with u_t (since $u_\varepsilon \to u$ uniformly) and thus u is a solution of (1.1).

Since u is also continuous up to $t = 0$ and $u(x, 0) = f(x)$, and since

$$|u| \leqq M \quad \text{by (4.1)},$$

u is uniquely determined [9; Chap. 2]. It follows that (4.6) (and (4.4)) hold for the full range of the parameter ε.

Consider now the parabolic equation

$$\begin{aligned} &u_t - u_{xx} = F(u), \\ &u(x, 0) = f(x), \end{aligned} \tag{4.8}$$

and set

$$N(t) \equiv \sup_{0<s<t} \sup_{x \in \mathbb{R}} u(x, t).$$

Then there exists a largest T_0 such that

$$N(t) < \infty \quad \forall\, t < T_0.$$

We assume that $T_0 < \infty$; T_0 is then called the *blow-up time* for (4.8).

LEMMA 4.3. *If the assumptions of Lemma* 4.1 *hold with* $T < T_0$, *then*

$$u_\varepsilon + |u_{\varepsilon,t}| \leqq A_T \quad \text{in } \mathbb{R} \times [0, T] \tag{4.9}$$

for all ε *sufficiently small, where*

$$A_T = \sup_{\mathbb{R} \times [0,T]} (u + |u_t|) + 1 \tag{4.10}$$

(*which is a positive constant independent of* M).

Proof. By Lemma 4.1, if ρ is sufficiently large then

$$u_\varepsilon(x, t) + |u_{\varepsilon,t}(x, t)| < 1 \quad \text{if } |x| > \rho,\ 0 \leqq t \leqq T.$$

On the other hand, if $|x| \leqq \rho$, $0 \leqq t \leqq T$ then, by (4.2) and Lemma 4.2, (4.9) holds provided ε is sufficiently small. □

5. lim inf $\phi_\varepsilon \geqq T_0$.

THEOREM 5.1. *Under the assumptions of Theorem* 2.1, *for any* $T_1 < T_0$, $\phi_\varepsilon(x) > T_1$ *for all* $x \in \mathbb{R}$ *provided* ε *is small enough, and*

$$u_\varepsilon \to u \text{ uniformly in } \mathbb{R} \times [0, T] \tag{5.1}$$

as $\varepsilon \to 0$.

Proof. From the proof of Theorem 2.1 we have that the conditions of Lemma 4.1 hold for some small T (see (3.2)). Lemma 4.3 thus implies that M in (4.1) can be replaced by the constant

$$A = \sup_{\mathbb{R}\times[0,T_1]} (u+|u_t|)+1,$$

provided ε is small enough.

Let $v(t)$ be the solution of

$$\begin{aligned} &\varepsilon v_{tt} + v_t = F(v), \\ &v(0) = A, \\ &v_t(0) = F(A). \end{aligned} \tag{5.2}$$

By (3.3), (3.4),

$$v(t) < 2A \quad \text{if } 0 < t \leq \sigma, \tag{5.3}$$

where σ is a positive constant independent of ε.

We wish to compare u_ε with $v(t-T+\delta)$ (for any $\delta > 0$) provided ε is sufficiently small (so that (4.9) is valid) in order to deduce that $u_\varepsilon(x, t)$ exist in $\mathbb{R}\times[0, \tilde{T}]$ for $\tilde{T} = T-\delta+\sigma$ (as long as $\tilde{T} \leqq T_1$), and

$$u_\varepsilon(x, t) \leqq \tilde{M} \quad \text{in } \mathbb{R}\times[0, \tilde{T}].$$

To do this we work with the solutions u^n of the truncated problems and proceed precisely as in the proof of Lemma 3.1, with $t=0$ replaced by $t = T-\delta$.

Since δ is arbitrary we deduce that the conditions of Lemma 4.1 hold with T replaced by $T+\sigma$. We can proceed in this way step-by-step until we reach the value $t = T_1$. □

COROLLARY 5.2. *Under the assumptions of Theorem* 2.1

$$\liminf_{\varepsilon\to 0} [\inf_x \phi_\varepsilon(x)] \geqq T_0. \tag{5.4}$$

6. lim sup $\phi_\varepsilon \leqq T_0$.

THEOREM 6.1. *Let the assumptions of Theorem* 2.1 *hold and assume also that* (1.14) *holds. Then, for any* $x \in \mathbb{R}$,

$$\limsup_{\varepsilon\to 0} \phi_\varepsilon(x) \leqq T_0. \tag{6.1}$$

Remark 6.1. Baras and Cohen [0] proved that if one approximates $F(u)$ in (1.1) by uniformly bounded smooth functions $F_n(u)$, then the corresponding solutions $u_n(x, t)$ converge to $+\infty$ for $t > T_0$. Theorem 6.1 is a somewhat analogous result for a different type of approximation, namely, for the solutions of (P_ε^1) as $\varepsilon \to 0$.

Proof. Suppose the assertion is not true. Then there exist $x_0 \in \mathbb{R}$ and $\delta > 0$ such that for a sequence $\varepsilon \to 0$,

$$\phi_\varepsilon(x_0) > T_0 + 2\delta.$$

By (2.6) we then get, for any $\rho > 0$,

$$\phi_\varepsilon(x) > T_0 + \delta \quad \forall x \in (-\rho, \rho), \tag{6.2}$$

provided ε is small enough.

From the definition of T_0 it follows that there is a sequence (x_n, η_n) with $\eta_n \downarrow 0$ such that

$$u(x_n, T_0 - \eta_n) \to \infty \quad \text{if } n \to \infty.$$

Choose any large positive constant M and let n_0 be such that

$$u_\varepsilon(x_{n_0}, T_0 - \eta_{n_0}) > M.$$

By Theorem 3.2, $u_\varepsilon(x, t)$ is monotone increasing in t and therefore

$$u_\varepsilon(x_{n_0}, t) > M \quad \text{if } T_0 - \eta_{n_0} \leqq t \leqq \phi_\varepsilon(x_{n_0}). \tag{6.3}$$

We choose ρ in (6.2) such that $\rho > |x_{n_0}| + 1$. Then

$$\phi_\varepsilon(x) > T_0 + \delta \quad \text{if } |x - x_{n_0}| \leqq 1. \tag{6.4}$$

Introduce the function

$$\psi(x) = \frac{\pi}{2} \sin \pi(x - x_{n_0}).$$

It satisfies

$$\begin{aligned} &\psi'' = -\pi^2 \psi, \quad \psi > 0 \quad \text{in } x_{n_0} < x < x_{n_0} + 1, \\ &\psi(x_{n_0}) = \psi(x_{n_0} + 1) = 0, \\ &\int_{x_{n_0}}^{x_{n_0}+1} \psi(x)\, dx = 1. \end{aligned} \tag{6.5}$$

Multiplying $\mathscr{L}_\varepsilon(u_\varepsilon) = F(u_\varepsilon)$ by ψ and integrating over $\{x_{n_0} < x < x_{n_0} + 1\}$, we find that the function

$$a(t) = \int_{x_{n_0}}^{x_{n_0+1}} u_\varepsilon(x, T_0 - \eta_{n_0} + t)\psi(x)\, dx$$

satisfies:

$$\varepsilon a'' + a' = -\pi^2 a + F(a) + [u_\varepsilon(x, T_0 - \eta_{n_0} + t)\psi_x(x)]_{x_{n_0}}^{x_{n_0+1}}.$$

Since

$$\psi_x(x_{n_0}) = -c < 0, \qquad \psi_x(x_{n_0} + 1) > 0,$$

it follows that

$$\varepsilon a'' + a' \geqq -\pi^2 a + F(a) + cM \tag{6.6}$$

where (6.3) was used; also

$$a(0) > 0, \quad a'(0) \geqq 0, \quad a'(t) \geqq 0. \tag{6.7}$$

□

LEMMA 6.2. *The solution $a(t)$ of* (6.6), (6.7) *blows up in time $t \leqq \delta$ provided M is sufficiently large and ε is sufficiently small.*

Assuming the lemma we conclude that

$$\phi_\varepsilon(x) < T_0 - \eta_{n_0} + \delta \quad \text{for some } x \in (x_{n_0}, x_{n_0+1}),$$

which is a contradiction to (6.4).

Proof of Lemma 6.2. Let $b(t)$ denote the solution of

$$\varepsilon b'' + b' = c_0(F(b) + M), \tag{6.8}$$

$$0 \leqq b(0) < a(0), \qquad b'(0) = a'(0) \tag{6.9}$$

where

$$c_0 = \min\left(\frac{1}{2}, \frac{c}{2}\right), \quad c \text{ as in (6.6).} \tag{6.10}$$

Writing (6.8) in the form

$$\varepsilon(e^{t/\varepsilon}b')' = c_0\, e^{t/\varepsilon}(F(b)+M) \geqq 0$$

we see that $b' \geqq 0$.

We claim that if M is large enough then, for any $\varepsilon > 0$,

$$b(t) \quad \text{blows up in time} \quad \leqq \delta. \tag{6.11}$$

Indeed, suppose $b(t)$ exists for all $t \leqq \delta$. We claim that there exists a t_0 such that

$$t_0 \in \left(0, \frac{\delta}{2}\right), \qquad b''(t_0) \geqq 0. \tag{6.12}$$

Indeed, otherwise we have

$$b''(t) < 0 \quad \forall t \in (0, \delta/2)$$

and therefore, by (6.8),

$$b' \geqq c_0(F(b)+M).$$

Hence

$$\int_0^{\delta/2} \frac{b'}{c_0(F(b)+M)} \geqq \frac{\delta}{2}.$$

But the left-hand side is bounded above by

$$\int_{a(0)}^{\infty} \frac{ds}{c_0(F(s)+M)}$$

which is $< \delta/2$ if M is sufficiently large; this is a contradiction.

Having proved (6.12), we differentiate (6.8) in t and obtain, after multiplying by $e^{t/\varepsilon}$,

$$\varepsilon(e^{t/\varepsilon}b'') = c_0\, e^{t/\varepsilon}F'(b)b' \geqq 0.$$

Using (6.12) we deduce that

$$b''(t) \geqq 0 \quad \text{if } t > t_0.$$

Hence

$$b'' + b' = c_0(F(b)+M) + (1-\varepsilon)b'' \geqq c_0(F(b)+M) \quad \text{if } t \geqq t_0.$$

Denoting by $\gamma(t)$ the solution of

$$\gamma'' + \gamma' = c_0(F(\gamma)+M), \quad \gamma(t_0) = b(t_0), \quad \gamma'(t_0) = b'(t_0), \tag{6.13}$$

we deduce that

$$\varepsilon(e^{t/\varepsilon}(b-\gamma)')' \geqq 0, \qquad (b-\gamma)(t_0) = (b-\gamma)'(t_0) = 0.$$

It follows that $b(t) \geqq \gamma(t)$. On the other hand, we can easily see that if M is sufficiently large then $\gamma(t)$ blows up in time $t \leqq t_0 + \gamma/2$. Therefore

$$b(t) \quad \text{blows up in time} \quad t < \delta. \tag{6.14}$$

To complete the proof of the lemma we compare $a(t)$ with $b(t)$. From (6.10) and (6.6), (6.8) we find that

$$\varepsilon(e^{t/\varepsilon}(a-b)')' \geqq c_0 F'(a-b)\, e^{t/\varepsilon}.$$

Also since $(a-b)(0)>0$, $(a-b)'(0)\geqq 0$, we easily deduce that $a(t)\geqq b(t)$ for all t for which $b(t)$ exists. It follows that $a(t)$ blows up in time $<\delta$.

7. Generalizations.

7.1. The results of this paper extend to the case where (1.7) is replaced by

$$g=f_{xx}+F(f)+\varepsilon h$$

provided h satisfies

$$|h(x)|\leqq \frac{C}{(1+|x|)^{\alpha}} \quad \text{for some } \alpha>0, \tag{7.1}$$

$$\sum_{i=1}^{3}|h^{(i)}(x)|\leqq C, \tag{7.2}$$

and

$$g\geqq 0. \tag{7.3}$$

These conditions ensure that (1.13) holds and that

$$\sum_{i=1}^{3}|D_x^i u_{tt}^{\varepsilon}(x,0)|\leqq M<\infty,$$

which is the only condition that $u_{tt}^{\varepsilon}(x,0)$ needed to satisfy in the previous analysis.

7.2. The results of this paper extend to the case where x is N-dimensional with $N=2$ or $N=3$, provided F, f and g satisfy the following additional conditions:

$$sF'(s)-F(s)\leqq 0 \quad \forall s\geqq 0, \tag{7.4}$$

$$\left(\frac{t}{2\varepsilon}+1\right)f(x)+tg(x)\geqq\frac{t}{\sqrt{\varepsilon}}|\nabla f(x)| \quad \forall t\geqq 0,\ x\in\mathbb{R}^N, \tag{7.5}$$

$$\begin{aligned}\frac{1}{2\sqrt{\varepsilon}}f(x)+\sqrt{\varepsilon}g(x)-\lambda|\nabla f(x)|+\frac{t}{\sqrt{\varepsilon}}\left[g(x)+\frac{1}{4\varepsilon}f(x)\right]-\lambda\left[\frac{1}{\sqrt{\varepsilon}}|\nabla f(x)|+\sqrt{\varepsilon}|\nabla g(x)|\right]\\ \geqq\frac{t}{\sqrt{\varepsilon}}|\nabla g(x)|+\lambda\frac{t}{\sqrt{\varepsilon}}|\nabla^2 f(x)| \quad \forall t\geqq 0,\ x\in\mathbb{R}^N\end{aligned} \tag{7.6}$$

where $\lambda>1$ (λ constant) (here $g=\Delta f+F(f)$). These conditions are satisfied if

$$F(s)=e^{\theta s}, \quad 0<\theta\leqq 1,$$

$$f(x)=A+f_1(x), \quad |D^{\alpha}f_1(x)|\leqq\frac{C}{(1+|x|)^{\beta}} \tag{7.7}$$

for $0\leqq|\alpha|\leqq 5$ and some $\beta>0$, and A is a sufficiently large positive constant.

Under these conditions the existence and uniqueness of a solution u_{ε} can be established using the formulation (P_{ε}^3) and the approximating sequence considered in [5]. Condition (7.5) ensures that $u_{\varepsilon}\geqq 0$. The existence of $\phi_{\varepsilon}(x)$ and (2.5), (2.6) result from the inequality

$$w_{\varepsilon,t}\geqq\lambda|\nabla w_{\varepsilon}|$$

which is proved using the extension of the representation formula (2.3) to N dimensions [5] and the conditions (7.4), (7.5). The results of §§ 4, 5 extend with minor changes to dimension N. Finally, in § 6 we need a stronger assumption than (1.14), namely,

$$g(x)\geqq\delta_0>0. \tag{7.8}$$

Using this we can prove as in [5] that $\exists \delta_1 > 0$ such that

$$u_{\varepsilon,t} \geqq \delta_1 |\nabla u_\varepsilon|.$$

This guarantees that, for any $t > T_0 + \sigma/2$ (σ arbitrarily small),

$$u_\varepsilon(x, t) > M \quad \text{for } x \text{ in a fixed ball } B \text{ of radius } \delta_1 \sigma/4.$$

Introducing the function

$$a(t) = \int_B u_\varepsilon\left(x, T_0 + \frac{\sigma}{4} + t\right) \psi(x)\, dx$$

where ψ is the principal eigenfunction of $-\Delta$ in B, $\psi > 0$ in B, $\int_B \psi = 1$, we again derive (6.6), (6.7), and conclude that $a(t)$ blows up in time $\leqq T_0 + \sigma$.

REFERENCES

[0] P. BARAS AND L. COHEN, *Complete blow-up after $T_{\max}$ for the solution of a semilinear heat equation*, J. Funct. Anal., 71 (1987), pp. 142–174.

[1] D. BOGY AND P. NAGHDI, *On heat conduction and wave propagation in rigid solids*, J. Math. Phys., 11 (1970), pp. 917–923.

[2] J. BREEZEL AND E. NOLAN, *Non-Fourier effects in the transmission of heat*, Proc. 6th Conf. on Thermal Conductivity, Dayton, Ohio, October 1966, pp. 237–254.

[3] J. BROWN, D. CHUNG, AND P. MATTHEWS, *Heat pulses at low temperatures*, Phys. Letters, 21 (1966), pp. 241–243.

[4] L. A. CAFFARELLI AND A. FRIEDMAN, *Differentiability of the blow-up curve for one-dimensional nonlinear wave equations*, Archive Rat. Mech. Anal., 91 (1985), pp. 83–98.

[5] ———, *The blow-up boundary for nonlinear wave equations*, Trans. Amer. Math. Soc., 297 (1986), pp. 223–241.

[6] C. CATTANEO, *Sulla concluzione del calore*, Atti Sem. Mat. Fis. Univ. Modena 3 (1948/49), pp. 3–21.

[7] M. CHESTER, *Second sound in solids*, Phys. Rev., 131 (1963), pp. 2013–2015.

[8] B. D. COLEMAN, M. FABRIZIO, AND D. R. OWEN, *Thermodynamics and the constitutive relations for second sounds in crystals*, in Lecture Notes in Physics 228, ed. H. Araki et al., Springer-Verlag, Berlin, New York, 1983, pp. 20–43.

[9] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

[10] ———, *Blow-up of solutions of nonlinear evolution equations*, in Directions in Partial Differential Equations, ed. M. Crandall et al., Academic Press, 1987, pp. 75–88.

[11] ———, *Blow-up of solutions of nonlinear parabolic equations*, in Nonlinear Diffusion Equations and their Equilibrium States, MSRI Publications, 12, Springer-Verlag, Berlin, New York, 1988, pp. 301–318.

[12] A. FRIEDMAN AND A. A. LACEY, *The blow up time for solutions of nonlinear heat equations with small diffusions*, SIAM J. Math. Anal., 18 (1987), pp. 711–721.

[13] A. FRIEDMAN AND B. MCLEOD, *Blow-up of positive solutions of semilinear parabolic equations*, Indiana Univ. Math. J., 34 (1985), pp. 425–447.

[14] A. FRIEDMAN AND L. OSWALD, *The blow-up surface of nonlinear wave equations with small spatial velocity*, Trans. Amer. Math. Soc., to appear.

[15] ———, *The blow-up time for higher order semilinear parabolic equations with small leading coefficients*, J. Differential Equations, to appear.

[16] Y. GIGA AND R. V. KOHN, *Asymptotically self-similar blow up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.

[17] ———, *Characterizing blow up using similarity variables*, Comm. Pure Appl. Math., 36 (1987), pp. 1–40.

[18] J. C. MAXWELL, Phil. Trans. Roy. Soc. London, 157 (1967), pp. 49–88.

[19] P. VERNOTTE, *Les paradoxes de la theorie continue d'equation de la chaleur*, Comp. Rend., 246 (1958), pp. 3154–3155.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 367-371, March 1989

005

A STRONG MAXIMUM PRINCIPLE FOR A NONCOOPERATIVE ELLIPTIC SYSTEM*

GUIDO SWEERS†

Abstract. In this note it is shown that on a ball in $\mathbb{R}^N$, with $N>2$, a maximum principle holds for a special elliptic system. This system is such that the classical maximum principle is not applicable.

Key words. maximum principle, positive solutions, elliptic systems, Green's function

AMS(MOS) subject classifications. 35B50, 35J55

Introduction and results. A linear elliptic system

$$-\Delta u_\alpha + \sum_{\beta=1}^{k} h_{\alpha\beta} u_\beta = f_\alpha \quad \text{with } \alpha \in \{1, 2, \cdots, k\} \tag{1}$$

is called cooperative (see [3]) if

$$h_{\alpha\beta} \leqq 0 \quad \text{for } \alpha \neq \beta. \tag{2}$$

If, moreover,

$$\sum_{\beta=1}^{k} h_{\alpha\beta} \geqq 0 \quad \text{for } \alpha \in \{1, 2, \cdots, k\} \tag{3}$$

we can extend the results of the maximum principle to system (1) (see [4, p. 191]).

The motivation for this note was the question of whether or not the cooperative property is necessary for obtaining a maximum principle. Recent results for noncooperative systems have been obtained by De Figueiredo and Mitidieri [1], and Weinberger [5].

We consider an elliptic system, with Dirichlet boundary conditions, which is in some sense the simplest noncooperative system:

$$\begin{aligned} -\Delta u &= f_1 - \lambda v \quad \text{in } \Omega, \\ -\Delta v &= f_2 \quad \text{in } \Omega, \\ u &= v = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4}$$

where Ω is the unit ball in $\mathbb{R}^N$, $N>2$. The classical maximum principle gives the following positivity result, which is not uniform.

LEMMA 1. *Let $f_1, f_2 \in C^\gamma(\bar{\Omega})$ for some $\gamma>0$ with $f_1 \geqq 0$ and $f_1 \neq 0$. Then there is $\lambda(f_1, f_2)>0$ such that for all $\lambda \in [0, \lambda(f_1, f_2))$ the solution u_λ of (4) satisfies*

$$u_\lambda > 0 \quad \text{in } \Omega. \tag{5}$$

The classical maximum principle does not show that it is possible to find a uniform result for $f_2 \leqq f_1$. Nevertheless, we prove for

$$\begin{aligned} -\Delta u &= f - \lambda v \quad \text{in } \Omega, \\ -\Delta v &= f \quad \text{in } \Omega, \\ u &= v = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{6}$$

the following theorem. (The same result holds for (4) if $f_2 \leqq f_1$ and $f_1 = f$.)

* Received by the editors November 30, 1987; accepted for publication (in revised form) June 15, 1988.
† Department of Mathematics and Informatics, Delft University of Technology, 2600 AJ Delft, the Netherlands.

THEOREM 2. (a) *There is a largest $\lambda_0 > 0$ for which the following holds. For all $(u, v) \in C^2(\bar{\Omega}) \times C^2(\bar{\Omega})$ satisfying* (6), *such that $\lambda < \lambda_0$, $f \geqq 0$ and $f \neq 0$, we find that*

$$u > 0 \quad \text{in } \Omega. \tag{7}$$

(b) $\lambda_0 \leqq (\lambda_1^{-1} + \lambda_2^{-1})^{-1} (< \lambda_1)$, *where λ_1, λ_2 are, respectively, the first and second eigenvalue of*

$$\begin{aligned} -\Delta\varphi &= \lambda\varphi \quad \text{in } \Omega, \\ \varphi &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{8}$$

Remark 1. If $N = 1$ then

$$u(x) = \int_{-1}^{1} \frac{1}{2}(1 - |x-y| - xy)\left(1 - \frac{\lambda}{6}(2 + 2|x-y| - x^2 - y^2)\right) f(y)\, dy. \tag{9}$$

Since $\max\{\frac{1}{6}(2 + 2|x-y| - x^2 - y^2); -1 \leqq x, y \leqq 1\} = \frac{2}{3}$ we find that $\lambda_0 = 3/2 \leqq (\lambda_1^{-1} + \lambda_2^{-1})^{-1} = \pi^2/5 \approx 1.97$. A direct calculation shows that $(\lambda_1^{-1} + \lambda_2^{-1} + \lambda_3^{-1} + \cdots)^{-1} = 3/2$. I cannot explain this similarity.

Remark 2. Let H be a subspace of $C(\bar{\Omega})$ such that the inverse B of $-\Delta$, with zero Dirichlet boundary condition, into $C^2(\bar{\Omega})$ is well defined. Theorem 2(a) then shows that

$$B(I - \lambda B)f > 0 \quad \text{in } \Omega \tag{10}$$

for all $\lambda < \lambda_0$ and $f \in H$ with $f \geqq 0$, $f \neq 0$.

Remark 3. The classical maximum principle [4, Thm. 2.2] shows that $Bf > 0$ for f as in (10). If also $f(x_0) = 0$ for some $x_0 \in \Omega$, then

$$((I - \lambda B)f)(x_0) < 0 \quad \text{for all } \lambda > 0. \tag{11}$$

Remark 4. Consider the system

$$\begin{aligned} -\Delta u &= f - \lambda^2 v \quad \text{in } \Omega, \\ -\Delta v &= u \quad \text{in } \Omega, \\ u &= v = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Hence for B as in Remark 2 and $\lambda < \lambda_1$

$$u = \left(\sum_{k=0}^{\infty} (\lambda^4 B^4)^k\right)(I + \lambda B)(I - \lambda B)Bf.$$

If $\lambda < \lambda_0$ as well, then Theorem 2(a) together with the classical maximum principle shows that if $f \geqq 0, f \neq 0$ then $u > 0$ in Ω. Both this system and (6) cannot be uncoupled as in [1, Remark 1.7] to find a maximum principle. Recent results concerning [1, Remark 1.7] can be found in [5].

Remark 5. Let Ω be an arbitrary domain, and let φ_1, φ_2 be the first and second eigenfunctions of (8), respectively. Set $H = \{c_1\varphi_1 + c_2\varphi_2;\ c_1, c_2 \in \mathbb{R}\}$. We can prove that $B(I - \lambda B)$ from H into H is positive if and only if $\lambda \leqq \lambda_0 = (\lambda_1^{-1} + \lambda_2^{-1})^{-1}$. We can also hope that in general $\lambda_0 = (\sum \lambda_i^{-1})^{-1}$, with the summation over all eigenfunctions. However, direct but tedious computations show that with $\Omega = (0, 1)$ and $H = \{c_1\varphi_1 + c_2\varphi_2 + c_3\varphi_3;\ c_1 \in \mathbb{R}\}$ the following inequality holds:

$$\lambda_0 < (\lambda_1^{-1} + \lambda_2^{-1} + \lambda_3^{-1})^{-1}.$$

Proofs. Lemma 1 can be proved by a straightforward application of the classical strong maximum principle. Let φ_1 be the first eigenfunction of (8) with $\varphi_1 > 0$ in Ω. Since $v = 0$ on $\partial\Omega$ and $v \in C^2(\bar{\Omega})$, there is $c_1 > 0$ such that

$$v \leqq c_1 \varphi_1 \quad \text{in } \Omega. \tag{12}$$

Let $w \in C^2(\bar{\Omega})$ be the solution of

$$\begin{aligned} -\Delta w &= v \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{13}$$

The maximum principle, [4, Thm. 2.6], then shows that

$$w \leqq \frac{c_1}{\lambda_1} \varphi_1 \quad \text{in } \Omega. \tag{14}$$

Since $-\Delta(u + \lambda w) = f_1 \geqq 0$ and $f_1 \not\equiv 0$, the strong maximum principle [4, Thms. 2.6, 2.7] implies that $u + \lambda w > c_2 \varphi_1$ in Ω for some $c_2 > 0$. Hence,

$$u > c_2 \varphi_1 - \lambda w \geqq c_2 \varphi_1 - \lambda \frac{c_1}{\lambda_1} \varphi_1 \geqq 0 \quad \text{in } \Omega \text{ if } \lambda \leqq \lambda_1 \frac{c_2}{c_1}. \tag{15}$$

□

Proof of Theorem 2(a, b). Equations (6) can be rewritten as

$$\begin{aligned} u(x) &= \int_\Omega G(x, y)(f(y) - \lambda v(y))\, dy \\ &= \int_\Omega G(x, y)\left(f(y) - \lambda \int_\Omega G(y, z) f(z)\, dz\right) dy \\ &= \int_\Omega \left(G(x, y) - \lambda \int_\Omega G(x, z) G(z, y)\, dz\right) f(y)\, dy, \end{aligned} \tag{16}$$

where (see [2, eqs. (2.12), (2.13)]),

$$\begin{aligned} G(x, y) &= g_n(|x - y|^{2-n} - |(|y|x - |y|^{-1}y)|^{2-n}), \qquad y \neq 0, \\ G(x, 0) &= g_n(|x|^{2-n} - 1). \end{aligned} \tag{17}$$

The Euclidean norm is denoted by $|\cdot|$, and $g_n = (n(n-2)\omega_n)^{-1}$, where ω_n is the volume of the unit ball in $\mathbb{R}^N$.

To prove the theorem, it is sufficient to show that

$$M(x, y) := (G(x, y))^{-1} \int_\Omega G(x, z) G(z, y)\, dz < M \tag{18}$$

for some $M < \infty$. We then find that $u > 0$ for all $\lambda \leqq \lambda_0 = M^{-1}$.

We will prove (18) by direct computations.

To simplify the notations we set

$$(xy) = |x - y|, \qquad (XY) = |(|y|x - |y|^{-1}y)|, \text{ etc.}$$

Note that $(XY)=(|y|^2|x|^2-2(x,y)+1)^{1/2}=(YX)$, and hence

$$(XY)^2-(xy)^2=|y|^2|x|^2+1-|x|^2-|y|^2 \tag{19}$$
$$=(1-|x|^2)(1-|y|^2)>0 \quad \text{for } x,y\in\Omega,$$

$$(xy)^{-1}-(XY)^{-1}=\frac{(1-|x|^2)(1-|y|^2)}{(XY)+(xy)}(xy)^{-1}(XY)^{-1} \quad \text{for } x\neq y. \tag{20}$$

Using (17)-(20) we find that

$$g_n^{-1}M(x,y)=\int_\Omega \frac{((xz)^{2-n}-(XZ)^{2-n})((yz)^{2-n}-(YZ)^{2-n})}{(xy)^{2-n}-(XY)^{2-n}}\,dz$$
$$=\int_\Omega \frac{((xz)^{-1}-(XZ)^{-1}\sum_{k=0}^{n-3}(xz)^{3-n+k}(XZ)^{-k}}{((xy)^{-1}-(XY)^{-1})\sum_{k=0}^{n-3}(xy)^{3-n+k}(XY)^{-k}}$$
$$\cdot((yz)^{-1}-(YZ)^{-1})\sum_{k=0}^{n-3}(yz)^{3-n+k}(YZ)^{-k}\,dz$$
$$=\int_\Omega \frac{(1-|x|^2)(1-|z|^2)}{1+(xz)(XZ)^{-1}}\,\frac{1+(xy)(XY)^{-1}}{(1-|x|^2)(1-|y|^2)}\,\frac{(1-|y|^2)(1-|z|^2)}{1+(yz)(YZ)^{-1}}$$
$$\cdot\frac{(\sum_{k=0}^{n-3}(xz)^{2-n+k}(XZ)^{-2-k})(\sum_{k=0}^{n-3}(yz)^{2-n+k}(YZ)^{-2-k})}{\sum_{k=0}^{n-3}(xy)^{2-n+k}(XY)^{-2-k}}\,dz \tag{21}$$
$$=(xy)^{n-2}(XY)^2\left(\sum_{k=0}^{n-3}\left(\frac{(xy)}{(XY)}\right)^k\right)^{-1}\int_\Omega \frac{(1-|z|^2)^2(1+(xy)(XY)^{-1})}{(1+(xz)(XZ)^{-1})(1+(yz)(YZ)^{-1})}$$
$$\cdot\left(\sum_{k=0}^{n-3}\left(\frac{(xz)}{(XZ)}\right)^k\right)\left(\sum_{k=0}^{n-3}\left(\frac{(yz)}{(YZ)}\right)^k\right)(yz)^{2-n}(xz)^{2-n}(YZ)^{-2}(XZ)^{-2}\,dz$$
$$\leqq(xy)^{n-2}(XY)^2\int_\Omega(1-|z|^2)^2 2(n-2)^2(yz)^{2-n}(xz)^{2-n}(YZ)^{-2}(XZ)^{-2}\,dz.$$

Assume $x\neq y$ and split Ω in Ω_1 and Ω_2, where

$$\Omega_1=\{z\in\Omega;\,|x-z|<|y-z|\},\qquad \Omega_2=\{z\in\Omega;\,|x-z|>|y-z|\}.$$

If $|x-z|<|y-z|$, then $|y-z|\geqq|y-x|-|x-z|\geqq|y-x|-|y-z|$ and

$$|(|y|x-|y|^{-1}y)|\leqq|y||x-z|+|(|y|z-|y|^{-1}y)|$$
$$\leqq|y-z|+|(|y|z-|y|^{-1}y)|.$$

Hence

$$(xy)\leqq 2(yz), \quad \text{and} \tag{22}$$

$$(XY)\leqq 2(YZ) \quad \text{for } z\in\Omega_1. \tag{23}$$

By exchanging x and y, respectively X and Y, we find equivalent inequalities for $z\in\Omega_2$. Moreover,

$$(XZ)^2=|x|^2|z|^2-2(x,z)+1$$
$$\geqq|x|^2|z|^2-2|x||z|+1$$
$$=(1-|x||z|)^2$$
$$\geqq(1-|z|)^2\geqq\tfrac{1}{4}(1-|z|^2)^2 \quad \text{for } z\in\Omega. \tag{24}$$

Combining (21)-(24) yields

$$
\begin{aligned}
M(x,y) &\leqq 2(n-2)^2 g_n \left(\int_{\Omega_1} \left(\frac{1-|z|^2}{(XZ)} \right)^2 \left(\frac{(xy)}{(yz)} \right)^{n-2} \left(\frac{(XY)}{(YZ)} \right)^2 (xz)^{2-n}\, dz \right. \\
&\qquad \left. + \int_{\Omega_2} \left(\frac{1-|z|^2}{(YZ)} \right)^2 \left(\frac{(xy)}{(xz)} \right)^{n-2} \left(\frac{(XY)}{(XZ)} \right)^2 (yz)^{2-n}\, dz \right) \\
&\leqq 2^{n+3}(n-2)^2 g_n \left(\int_{\Omega_1} (xz)^{2-n}\, dz + \int_{\Omega_2} (yz)^{2-n}\, dz \right) \\
&< 2^{n+4}(n-2)^2 g_n \int_{2\Omega} |z|^{2-n}\, dz = 2^{n+5} \frac{(n-2)\omega_{n-1}}{n\omega_n},
\end{aligned} \tag{25}
$$

which completes the proof of Theorem 2(a).

To prove Theorem 2(b) let φ_1 and φ_2 be, respectively, the first and second eigenfunctions of (8), with $\varphi_1 > 0$ in Ω.

First note that $\lambda_0 \leqq \lambda_1$. Indeed, if $\lambda > \lambda_1$ then $u = \lambda_1^{-1}(1-\lambda/\lambda_1)\varphi_1$ is a solution of (6) with $f = \varphi_1 > 0$ in Ω, while $u < 0$ in Ω. Suppose $(\lambda_1^{-1}+\lambda_2^{-1})^{-1} < \lambda \leqq \lambda_1$ and let $c > 0$ be the largest constant such that

$$
\varphi_1 - c\varphi_2 \geqq 0 \quad \text{in } \Omega. \tag{26}
$$

Let U be the solution of (6) with $f = \varphi_1 - c\varphi_2$. Then

$$
U = \lambda_1^{-2}(\lambda_1 - \lambda)(\varphi_1 - c\varphi_2) - c(\lambda_1^{-2} - \lambda_2^{-2})(\lambda - (\lambda_1^{-1}+\lambda_2^{-1})^{-1})\varphi_2. \tag{27}
$$

If $\lambda = \lambda_1$ then U is negative somewhere in Ω since φ_2 changes sign. If $\lambda < \lambda_1$ then U is negative somewhere since c is the largest constant such that (26) holds, which is a contradiction. □

Acknowledgment. I thank E. Mitidieri for many stimulating and helpful discussions.

REFERENCES

[1] D. G. De Figueiredo and E. Mitidieri, *A maximum principle for an elliptic system and applications to semilinear problems*, SIAM J. Math. Anal., 17 (1986), pp. 836-849.

[2] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.

[3] M. W. Hirsch, *Systems of differential equations which are competitive or cooperative*, I: *limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167-179.

[4] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[5] H. F. Weinberger, *Some remarks on invariant sets for systems*, in Proc. Conference on Maximum Principles and Eigenvalue Problems in Partial Differential Equations, P. Schaefer, ed., Research Notes, University of Tennessee, Longman Press, London, 1987.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 372-387, March 1989

006

A SINGULAR PERTURBATION ANALYSIS OF REVERSE-BIASED SEMICONDUCTOR DIODES*

F. BREZZI†, A. C. S. CAPELO‡, AND L. GASTALDI§

Abstract. The one-dimensional equations of semiconductor devices are presented and studied here as a singular perturbation problem. It is shown that the limit problem, for reverse-biased devices, is a variational inequality of double obstacle type, which justifies, in some sense, the so-called total depletion assumption often used by engineers. It is also proven that the solution is unique if the value of the singular perturbation parameter is small enough.

Key words. semiconductors, singular perturbations

AMS(MOS) subject classifications. 78A30, 34B15, 34C11, 34E15

1. Introduction. The basic equations governing carrier transport in semiconductor devices are [10]:

$$\operatorname{div}(\varepsilon\nabla\psi) = -q(D-n+p) \quad \text{in } \Omega\times[0,T], \tag{1.1}$$

$$\frac{\partial p}{\partial t} = \operatorname{div}(D_p\nabla p + \mu_p p\nabla\psi) + R \quad \text{in } \Omega\times[0,T], \tag{1.2}$$

$$\frac{\partial n}{\partial t} = \operatorname{div}(D_n\nabla n - \mu_n n\nabla\psi) + R \quad \text{in } \Omega\times[0,T], \tag{1.3}$$

(1.4) suitable initial and boundary conditions,

where $\psi(t,x)$, the electric potential; $p(t,x)$, the concentration of positively charged holes; and $n(t,x)$, the concentration of negatively charged conduction electrons, are the unknowns. The function $D(x)$, the doping profile, is supposed to be known, as well as $\varepsilon(x)$, the permitivity of the semiconductor material. On the other hand, $\mu_p(\nabla\psi)$ and $\mu_n(\nabla\psi)$, the hole and electron mobilities, $D_p(\nabla\psi)$ and $D_n(\nabla\psi)$, the hole and the electron diffusion coefficients, and the generation-recombination term $R(p,n,\nabla\psi)$ are supposed to be known functions of $\nabla\psi$ and $p, n, \nabla\psi$, respectively. Finally q is the charge of the electron.

In the present paper we study a simplified version of (1.1)-(1.4). In particular:

(1) We consider a one-dimensional case: $\Omega =]-a, b[$ with $a, b > 0$.

(2) We assume ε, μ_p, μ_n, D_p, D_n to be constants.

(3) We consider the stationary case: $\partial p/\partial t = \partial n/\partial t = 0$.

(4) We neglect the generation-recombination term: $R \equiv 0$; this makes sense at least for reverse-biased devices, such as those we will consider.

(5) We assume $D(x)$ to be of the form

$$D(x) = D_0 \operatorname{sign} x, \qquad (D_0 > 0) \tag{1.5}$$

(in particular this implies that the junction is at $x=0$); in fact, to assume that $D(x) = D_1 < 0$ for $x < 0$ and $D(x) = D_2 > 0$ for $x > 0$ (which is far more realistic) would only complicate the notation, and would not change the results.

* Received by the editors March 2, 1987; accepted for publication (in revised form) May 13, 1988. This work was supported in part by the M.P.I.

† Dipartimento di Meccanica Strutturale della Università di Pavia and Istituto di Analisi Numerica del C.N.R., Pavia, Italy.

‡ Dipartimento di Scienze Statistiche dell'Università di Padova, Padova, Italy.

§ Dipartimento di Matematica dell'Università di Trento, Trento, Italy.

With assumptions (1)-(5), (1.1)-(1.3) can now be rewritten, after a suitable scaling, as follows:

$$\psi_s'' = -D_s + n_s - p_s \quad \text{in }]-a, b[, \qquad (\psi_s = \varepsilon\psi/qD_0;\ D_s = D/D_0), \tag{1.6}$$

$$(\lambda p_s' + p_s\psi_s')' = 0 \quad \text{in }]-a, b[, \qquad (p_s = p/D_0), \tag{1.7}$$

$$(\lambda n_s' - n_s\psi_s')' = 0 \quad \text{in }]-a, b[, \qquad (n_s = n/D_0). \tag{1.8}$$

Note that the same λ appears in (1.7) and (1.8) because we assume Einstein's relations

$$D_p/\mu_p = D_n/\mu_n = kT/q \tag{1.9}$$

hold (here k is the Boltzmann constant and T is the absolute temperature). The dependence of λ on the other physical quantities is given by

$$\lambda = \varepsilon kT/q^2 D_0, \tag{1.10}$$

which shows that $[\lambda]$ = square meters. For a piece of silicon at room temperature and a doping $D_0 = 10^{23}/\text{m}^3$, (1.10) gives $\lambda \simeq 10^{-16}\,\text{m}^2$. Let d be the actual length of the device (a typical value would be $d = 10^{-5}$ m). By scaling the domain to one, we obtain $\lambda/d^2 \simeq 10^{-16}\,(\text{m}/d)^2$, which is now a good dimensionless "stiffness parameter." From now on we will assume that d is used as the unit for length. In particular, $b + a = 1d$. Hence, a typical value for λ would be $\lambda \simeq 10^{-6}\,d^2$. Note that equations (1.6)-(1.8) are not dimensionless. In particular, $[\psi_s] = [d^2] = L^2$, $[p_s] = [1]$, $[n_s] = [1]$, and $[x] = [d] = L$.

It is natural to ask what is the limit of the solutions of (1.6)-(1.8) (with suitable boundary conditions) as λ/d^2 goes to zero. This has been done previously [2]-[4] with some additional simplifications. In particular, in [2] and [4] the unipolar case (say $D_s \equiv -1$ and $n_s \equiv 0$) has been studied for one-dimensional and two-dimensional domains (respectively), while in [3] the full set of three equations has been considered (in one dimension) but with simplified boundary conditions. In the present paper we consider the limit of (1.6)-(1.8) for λ/d^2 going to zero with more realistic boundary conditions. More precisely, the physical boundary conditions are given (before scaling) by

$$p(-a) = D_0(1+\beta), \qquad p(b) = D_0\beta, \tag{1.11}$$

$$n(-a) = D_0\beta, \qquad n(b) = D_0(1+\beta), \tag{1.12}$$

$$\psi(-a) = -V - \frac{kT}{2q}\log\frac{1+\beta}{\beta}, \qquad \psi(b) = V + \frac{kT}{2q}\log\frac{1+\beta}{\beta}. \tag{1.13}$$

Here the externally applied potential is assumed to be $-V$ *at* $x = -a$ and V at $x = b$ (reverse bias). In (1.11)-(1.13) β is the solution of

$$\beta(1+\beta) = \nu_i^2 = n_i^2/D_0^2 \tag{1.14}$$

with

$$n_i = 2(2\pi kT/h^2)^{3/2}(m_n m_p)^{3/4}\, e^{-W_f/2kT} \tag{1.15}$$

(h is Planck's constant, m_n and m_p are the effective mass of an electron and hole, respectively, and W_f is the width of the semiconductor forbidden band). After scaling, (1.11)-(1.13) become

$$p_s(-a) = 1+\beta, \qquad p_s(b) = \beta, \tag{1.16}$$

$$n_s(-a) = \beta, \qquad n_s(b) = 1+\beta, \tag{1.17}$$

$$\psi_s(-a) = -\bar{\psi} = -\alpha - \gamma, \qquad \psi_s(b) = \bar{\psi} = \alpha + \gamma, \tag{1.18}$$

with

$$\alpha = \frac{\varepsilon}{qD_0} V, \qquad \gamma = \frac{\lambda}{2} \log \frac{1+\beta}{\beta}. \tag{1.19}$$

In [3] the simplified boundary conditions

$$p_s(-a) = 1, \qquad p_s(b) = 0, \tag{1.20}$$

$$n_s(-a) = 0, \qquad n_s(b) = 1, \tag{1.21}$$

$$\psi_s(-a) = -\alpha, \qquad \psi_s(b) = \alpha \tag{1.22}$$

were assumed instead of (1.16)-(1.18).

If we want to study the limit of the problem (1.6)-(1.8) plus (1.16)-(1.18) for λ going to zero, we should express the dependence on λ of the boundary conditions (1.16)-(1.18) as well. Let us look again at (1.10). Since it is not reasonable to assume that k, ε, or q can change, we may think either the temperature T goes to zero or the doping D_0 goes to infinity. At first sight perhaps the second possibility looks physically more appealing, but it has the drawback that D_0 also enters in the scaling so that in the end we would not clearly understand the significance of the limit problem for the original unscaled variables. The first possibility does not have this drawback, and therefore we choose to let the *temperature* go to zero both in (1.10) (as was done first in [2] and then in [1], [3]) and in (1.15). Still we point out explicitly that the limit for $T \to 0$ has to be understood as just a mathematical "trick" in order to have better insight into the qualitative behavior of the solutions for the *physical* value of λ (or, rather λ / d^2), which is actually very small.

If we assume that the variable parameter in (1.10) is the temperature T, we can then express the behavior of the boundary conditions (1.16)-(1.18) as a function of λ as follows:

$$\beta = \tfrac{1}{2}(\sqrt{1+4B\lambda^3 e^{-A/\lambda}} - 1), \tag{1.23}$$

$$\gamma = \frac{\lambda}{2} \log (\sqrt{1+4B\lambda^3 e^{-A/\lambda}} + 1)/(\sqrt{1+4B\lambda^3 e^{-A/\lambda}} - 1), \tag{1.24}$$

with

$$A = W_f \varepsilon / q^2 D_0, \tag{1.25}$$

$$B = 4D_0(2\pi q^2/h^2\varepsilon)^3 (m_n m_p)^{3/2}. \tag{1.26}$$

Note that A is the scaled value of the potential difference associated with the energy gap between valence and conduction band, which is a kind of built-in potential at 0°K.

In the present paper we study the limit for $\lambda \to 0$ of (1.6)-(1.8) with the boundary conditions (1.16)-(1.18) and the relations (1.23)-(1.24). We show that as $\lambda \to 0$ the solutions of this problem tend to the solution of the double obstacle problem: find $\psi_0 \in C^1([-a, b])$ such that

$$\psi_0(-a) = -(\alpha + A/2), \qquad \psi_0(b) = \alpha + A/2, \tag{1.27}$$

$$-(\alpha + A/2) \leqq \psi_0 \leqq \alpha + A/2 \quad \text{in }]-a, b[, \tag{1.28}$$

$$\psi_0'' \geqq -D_s(x) \quad \text{when } \psi_0 > -(\alpha + A/2), \tag{1.29}$$

$$\psi_0'' \leqq -D_s(x) \quad \text{when } \psi_0 < \alpha + A/2. \tag{1.30}$$

Moreover, p_0 and n_0 are then the characteristic functions of the sets $\{\psi_0 = -\alpha - A/2\}$ and $\{\psi_0 = \alpha + A/2\}$, respectively. In some sense this justifies the so-called "total depletion assumption," which is often used for reverse-biased devices.

We also show that the solution of (1.6)-(1.8) plus (1.16)-(1.18) is unique for λ small enough. To our knowledge, three techniques are available for proving the uniqueness of the solution of this problem.

(1) For λ big enough, the nonlinearity of (1.7)-(1.8) becomes negligible and the uniqueness of the solution is trivial. However, this case is physically unreasonable.

(2) For α small enough we have a small perturbation of the nonlinear problem $\psi_s'' = -D_s(x) - \nu_i e^{-\psi_s/\lambda} + \nu_i e^{\psi_s/\lambda}$ that is of monotone type. This approach is followed, for instance, in [9]. See also [7] for additional references.

(3) For λ small enough, we can take advantage of the uniqueness of the limit for $\lambda \to 0$ of the solutions of the problem itself. This approach has been followed in [3] for the simplified boundary conditions.

Here the result is proved for the physical boundary conditions. However, we point out that uniqueness is not expected to hold in more general two-dimensional cases.

Finally, we remark that, although the difference between the simplified boundary conditions (1.20)-(1.22) and the physical ones (1.16)-(1.18) is very small (of order 10^{-10} or less for silicon at room temperature, for example), the qualitative behavior of the solutions changes. For instance, with the simplified boundary conditions we have, in general, $\psi_s'(-a) < 0$ and $\psi_s'(b) < 0$, while with the physical boundary conditions we have $\psi_s'(x) > 0$ everywhere. Similarly, we have $J_p := (\lambda p_s' + p_s \psi_s') < 0$ with the simplified boundary conditions and $J_p > 0$ with the "true" boundary conditions.

For a different scaling and a different asymptotic analysis see, for instance, [6], [7], and the references therein. For an analysis of the present scaling in some time-dependent problems, see [8].

2. An existence theorem. For the reader's convenience we rewrite the full problem in its scaled version here. We delete the subscript s.

PROBLEM 1. Given the positive constants a, b, λ, α, A, and B, find functions ψ, p, and n such that

$$\psi'' = -\operatorname{sign} x - p + n \quad \text{in }]-a, b[, \tag{2.1}$$

$$|(\lambda p' + p\psi')' = 0 \quad \text{in }]-a, b[, \tag{2.2}$$

$$(\lambda n' - n\psi')' = 0 \quad \text{in }]-a, b[, \tag{2.3}$$

$$\psi(-a) = -\bar{\psi}, \qquad \psi(b) = \bar{\psi}, \tag{2.4}$$

$$p(-a) = 1 + \beta, \qquad p(b) = \beta, \tag{2.5}$$

$$n(-a) = \beta, \qquad n(b) = 1 + \beta \tag{2.6}$$

where

$$\bar{\psi} = \alpha + \frac{\lambda}{2} \log \frac{1+\beta}{\beta}, \tag{2.7}$$

and β is the positive solution of

$$\beta(1+\beta) = B\lambda^3 e^{-A/\lambda} =: \nu_i^2. \tag{2.8}$$

In the following we use, as an abbreviation, the notation

$$D(x) = -\operatorname{sign} x. \tag{2.9}$$

We also introduce the so-called Slotboom variables:

$$\rho(x) = p(x) \exp[(\psi - \bar{\psi})/\lambda], \tag{2.10}$$

$$\sigma(x) = n(x) \exp[-(\bar{\psi} + \psi)/\lambda]. \tag{2.11}$$

We use them at once to show some interesting properties of the solutions of Problem 1 (if any exist). Substituting (2.10) and (2.11) into (2.2),(2.5) and (2.3),(2.6), respectively, we get

$$(\exp[-(\psi-\bar{\psi})/\lambda]\rho')'=0, \quad \rho(-a)=\beta\exp(-2\alpha/\lambda), \quad \rho(b)=\beta, \tag{2.12}$$

$$(\exp[(\psi+\bar{\psi})/\lambda]\sigma')'=0, \quad \sigma(-a)=\beta, \quad \sigma(b)=\beta\exp(-2\alpha/\lambda). \tag{2.13}$$

From this and the maximum principle we obtain

$$\beta\exp(-2\alpha/\lambda)\leqq\rho\leqq\beta, \tag{2.14}$$

$$\beta\exp(-2\alpha/\lambda)\leqq\sigma\leqq\beta, \tag{2.15}$$

and moreover it follows that ρ is a nondecreasing function and σ is a nonincreasing function. Finally, from (2.10)-(2.15) we obtain the well-known properties (see, e.g. [9], [11])

$$\beta(\beta+1)\exp(-2\alpha/\lambda)\leqq pn=\rho\sigma\exp(2\bar{\psi}/\lambda)\leqq\beta(\beta+1)\exp(2\alpha/\lambda) \tag{2.16}$$

and

$$p>0, \qquad n>0, \tag{2.17}$$

which are the properties we mentioned above. Let us remark also that solving (2.2) and (2.3) for p and n, using the boundary conditions (2.4)-(2.6), we get, respectively,

$$p(x)=\exp[(\bar{\psi}-\psi)/\lambda]\beta\left\{e^{-2\alpha/\lambda}\int_x^b e^{\psi/\lambda}\Big/\int_{-a}^b e^{\psi/\lambda}+\int_{-a}^x e^{\psi/\lambda}\Big/\int_{-a}^b e^{\psi/\lambda}\right\}, \tag{2.18}$$

$$n(x)=\exp[(\bar{\psi}+\psi)/\lambda]\beta\left\{e^{-2\alpha/\lambda}\int_{-a}^x e^{-\psi/\lambda}\Big/\int_{-a}^b e^{-\psi/\lambda}+\int_x^b e^{-\psi/\lambda}\Big/\int_{-a}^b e^{-\psi/\lambda}\right\} \tag{2.19}$$

and then we can write (2.2) and (2.3) in the form

$$\lambda p'+p\psi'=:C_p=\lambda\nu_i(e^{\alpha/\lambda}-e^{-\alpha/\lambda})\Big/\int_{-a}^b e^{\psi/\lambda}, \tag{2.20}$$

$$\lambda n'-n\psi'=:C_n=\lambda\nu_i(e^{-\alpha/\lambda}-e^{\alpha/\lambda})\Big/\int_{-a}^b e^{-\psi/\lambda}. \tag{2.21}$$

With these preliminaries we outline the proof of the following well-known result (see, for instance, [5], [7]).

THEOREM 1. *For every $\lambda>0$ there exists at least one solution $\{\psi, p, n\}$ of Problem* 1.

Proof. Let

$$\chi=(1+\sqrt{1+4\nu_i^2\exp(2\alpha/\lambda)})/2 \tag{2.22}$$

be the maximum root of the second-degree equation

$$\chi^2-\chi-\nu_i^2e^{2\alpha/\lambda}=0 \tag{2.23}$$

and consider the set

$$K_\chi=\{(p, n): 0\leqq p\leqq\chi, 0\leqq n\leqq\chi\}\subset(L^\infty(-a, b))^2. \tag{2.24}$$

Now we define on K_χ a mapping $T:(p, n)\to(\bar{p}, \bar{n})$ in the following way. For fixed $(p, n)\in K_\chi$ let Ψ be the solution of

$$\Psi''=D-p+n \quad \text{in }]-a, b[, \quad \Psi(-a)=-\bar{\psi}, \quad \Psi(b)=\bar{\psi}, \tag{2.25}$$

and then let P and N be the solution of

$$\lambda P''+P'\Psi'+P\Psi''=0 \quad \text{in }]-a, b[, \quad P(-a)=1+\beta, \quad P(b)=\beta, \tag{2.26}$$

$$\lambda N''-N'\Psi'-N\Psi''=0 \quad \text{in }]-a, b[, \quad N(-a)=\beta, \quad N(b)=1+\beta \tag{2.27}$$

(which are given explicitly by (2.18) and (2.19), respectively, with Ψ in the place of ψ). Now let us set

$$\bar{p}=\min(\chi, P), \qquad \bar{n}=\min(\chi, N). \tag{2.28}$$

This defines the operator T. Obviously, T maps K_χ into itself. On the other hand, T is compact and therefore has a fixed point. Let $(\tilde{p}, \tilde{n})$ be such a fixed point and $\tilde{\psi}$ the corresponding solution of (2.1). To prove that $(\tilde{p}, \tilde{n}, \tilde{\psi})$ is a solution of Problem 1, it is enough to show that the corresponding functions $\tilde{P}$, $\tilde{N}$ satisfy $\tilde{P}\leqq\chi$, $\tilde{N}\leqq\chi$. Suppose to the contrary that, e.g., $\tilde{P}$ has a maximum $\tilde{P}(\bar{x})>\chi$ at $\bar{x}\in]-a, b[$. Then $\tilde{P}''(\bar{x})\leqq 0$, $\tilde{P}'(\bar{x})=0$, and from (2.26) it follows that $\tilde{P}(\bar{x})\tilde{\Psi}''(\bar{x})\geqq 0$, so that $\tilde{\Psi}''(\bar{x})\geqq 0$. Now

$$\begin{aligned} 0\leqq\tilde{\Psi}''(\bar{x}) &= D(\bar{x})-\tilde{p}(\bar{x})+\tilde{n}(\bar{x}) \\ &\leqq 1-\chi+\tilde{N}(\bar{x})\leqq 1-\chi+\nu_i^2\, e^{2\alpha/\lambda}/\tilde{P}(\bar{x})<1-\chi+\nu_i^2\, e^{2\alpha/\lambda}/\chi \end{aligned} \tag{2.29}$$

(where we have used the fact that $\tilde{P}$ and $\tilde{N}$ satisfy (2.16)), and this contradicts the definition of χ. □

Finally, let us note, as follows from (2.1), (2.18), and (2.19), that any solution of Problem 1 is of class C^1 in $]-a, b[$ and analytic in $]-a, 0[$ and in $]0, b[$.

3. Qualitative properties of the solutions. In this section we will study some qualitative properties of the solutions of Problem 1. The objective is twofold. On the one hand, we want to know the behavior of p, n, and ψ, which is of interest in itself, and on the other hand, the results obtained in this section will be used to prove an important a priori estimate in the next section.

PROPOSITION 1. *Let $\bar{x}\in[-a, b]$, $\bar{x}\neq 0$. Then the following two assertions hold:*

(a) *If $p(\bar{x})-D(\bar{x})<n(\bar{x})$, then $p-D$ cannot have a minimum at $\bar{x}$ and n cannot have a maximum at $\bar{x}$.*

(b) *If $p(\bar{x})-D(\bar{x})>n(\bar{x})$, then $p-D$ cannot have a maximum at $\bar{x}$ and n cannot have a minimum at $\bar{x}$.*

Proof. The result is an easy consequence of the maximum principle applied to (2.2) and (2.3). □

Remark 1. Let $\bar{x}\in[-a, b]$. If $p(\bar{x})-D(\bar{x}-)<n(\bar{x})$ and $p'(\bar{x})=0$, then p cannot be decreasing in a left neighborhood of $\bar{x}$. To see this it is enough to calculate $\lim_{x\to\bar{x}-} p''(x)$. Mutatis mutandis we can state and prove several similar results (with n instead of p, $>$ instead of $<$ and $\bar{x}+$ instead of $\bar{x}-$). □

Now let us define the following two functions:

$$M(x)=\max(p-1, n) \quad \text{in } [-a, 0], \tag{3.1}$$

$$m(x)=\min(p-1, n), \quad \text{in } [-a, 0]. \tag{3.2}$$

PROPOSITION 2. *M has no maxima and m has no minima in $]-a, 0[$.*

Proof. Let us suppose that M has a maximum at $\bar{x}\in]-a, 0[$. From Proposition 1, it follows at once that $p(\bar{x})-1=n(\bar{x})$ and therefore $M(\bar{x})=m(\bar{x})$. On the other

hand, we will have also $M'(\bar{x}) = m'(\bar{x}) = 0$ and $p'(\bar{x}) = n'(\bar{x}) = 0$. From this and (2.1)-(2.3), we infer that $p''(\bar{x}) = n''(\bar{x}) = 0$. Now, taking successive derivatives of (2.1)-(2.3), we conclude that all the derivatives of p and n vanish at $\bar{x}$. Since p and n are analytic we have $p = 1+\beta$ and $n = \beta$ in $[-a, 0]$. Note that in this case we have $\psi''(0-) = 0$. On the other hand, since $p(x)$ and $n(x)$ are continuous at 0, (2.1) implies $\psi''(0+) = -2$, so that, from (2.2) and (2.3), $p''(0+) > 0$ and $n''(0+) < 0$. Then in $[0, b]$ $p - D$ is increasing near zero and n is decreasing near zero. Since $p - D = 2+\beta$ and $n = \beta$ at 0+ while $p - D = 1+\beta = n$ at b, we would contradict Proposition 1 in $[0, b]$. The same reasoning shows that m cannot have a minimum in $]-a, 0[$. □

PROPOSITION 3. *ψ satisfies $\psi'(-a) > 0$ and $\psi'(b) > 0$.*

Proof. The proofs of the two assertions are similar. Let us prove the first one. Assume to the contrary that $\psi'(-a) \leqq 0$. Equations (2.20) and (2.21) imply that $p'(-a) > 0$ and $n'(-a) < 0$. Then $M = p - 1$ and $m = n$ in $[-a, 0]$ and for every $x \in [-a, 0]$ we have $M'(x) > 0$ and $m'(x) < 0$ (otherwise Proposition 1 or Remark 1 would be contradicted). Hence $p - D > 2+\beta$ at 0+, with $p'(0) > 0$, $n'(0) < 0$, and $p(0) - 1 > n(0)$. This is impossible because of Proposition 1 and $p - D = 1+\beta$ at b. □

PROPOSITION 4. *If m is decreasing in $[-a, 0]$, then $z := n - p + 1 > 0$ in $]-a, 0]$.*

Proof. To the contrary assume first that z has a negative minimum at $\bar{x} \in]-a, 0[$. Hence we have $z(\bar{x}) < 0$, $z'(\bar{x}) = 0$, and $z''(\bar{x}) \geqq 0$. Since $z'(\bar{x}) = 0$ we have $n'(\bar{x}) = p'(\bar{x})$ and, since $m(x)$ is decreasing, $n'(\bar{x}) = p'(\bar{x}) \leqq 0$. Inserting this in (2.20) we get $\psi'(\bar{x}) > 0$. Now, subtracting (2.3) from (2.2), we get

$$-\lambda z'' + (p' + n')\psi' + (p + n)z = 0, \tag{3.3}$$

which is impossible. Now assume that z has a zero minimum at $\bar{y} \in]-a, 0[$. Hence we have $z(\bar{y}) = 0$, $z'(\bar{y}) = 0$, and $z''(\bar{y}) \geqq 0$, and therefore $p'(\bar{y}) = n'(\bar{y}) \leqq 0$ and $\psi'(\bar{y}) > 0$. Because of (3.3), $z''(\bar{y}) = 0$ and $p'(\bar{y}) = n'(\bar{y}) = 0$. This implies, according to the reasoning given in the proof of Proposition 2, that $p = 1+\beta$ and $n = \beta$ in $]-a, 0[$, which is impossible. Hence we have shown that z cannot have a nonpositive minimum in $]-a, 0[$. To conclude the proof we must show that $z(0) > 0$. Let us assume that this is not true, i.e., $z(0) \leqq 0$, that is $n(0) \leqq p(0) - 1$ and $z(0) < z(x)$ for all $x \in [-a, 0]$. This and the hypothesis on m imply that $n(0) < \beta$ and $n'(0) \leqq 0$, which contradicts Propositions 1 and 2 (or Remark 1) because $p(0+) - D(0+) > 2 > n(0+)$ and $n(b) = p(b) - D(b) = 1+\beta$. □

PROPOSITION 5. *If m is decreasing in $[-a, 0]$ then we have the following:*

(i) *$m = p - 1 < M = n$ in $]-a, 0]$;*

(ii) *$m = p - 1$ is strictly decreasing in $[-a, 0]$;*

(iii) *$M = n$ has at most one minimum (and here $n < \beta$) in $]-a, 0]$;*

(iv) *If there exists $\bar{x} \in]-a, 0]$ such that $n(\bar{x}) > \beta$, then $n'(\bar{x}) > 0$.*

Proof. Assertions (i) and (ii) follow from Proposition 4 and the hypothesis. Assertion (iii) follows from Proposition 2. Finally Proposition 1 (or Remark 1) gives (iv). □

Analogous results hold in $[0, b]$ for the functions

$$\bar{M}(x) = \max(p, n - 1) \quad \text{in } [0, b], \tag{3.4}$$

$$\bar{m}(x) = \min(p, n - 1) \quad \text{in } [0, b]. \tag{3.5}$$

More precisely we have the following propositions.

PROPOSITION 6. *If $\bar{m}$ is increasing in $[0, b]$, then $z := n - p - 1 < 0$ in $[0, b[$.*

PROPOSITION 7. *If $\bar{m}$ is increasing in $[0, b]$, then we have the following:*

(i) *$\bar{m} = n - 1 < \bar{M} = p$, in $[0, b[$;*

(ii) *$\bar{m} = n - 1$ is strictly increasing in $[0, b]$;*

(iii) *$\bar{M} = p$ has at most one minimum (and here $p < \beta$) in* $[0, b[$;

(iv) *If there exists $\bar{y} \in [0, b[$ such that $p(\bar{y}) > \beta$, then $p'(\bar{y}) < 0$.*

We now verify the hypotheses on m and $\bar{m}$.

PROPOSITION 8. *The function m is decreasing in* $[-a, 0]$.

Proof. If m is not decreasing, we must have $m'(-a) \geqq 0$ by Proposition 2. Hence $p'(-a) \geqq 0$ and $n'(-a) \geqq 0$. Now, adding (2.20) and (2.21), we get

$$\lambda(p' + n') = C_p + C_n + (n - p)\psi'. \tag{3.6}$$

This yields, for $x = -a$,

$$\lambda(p'(-a) + n'(-a)) = C_p + C_n - \psi'(-a) \geqq 0 \tag{3.7}$$

so that, by Proposition 3,

$$C_p + C_n > 0. \tag{3.8}$$

On the other hand, (3.6) for $x = b$ yields

$$\lambda(p'(b) + n'(b)) = C_p + C_n + \psi'(b) > 0, \tag{3.9}$$

from which it follows that $p'(b) > 0$ and/or $n'(b) > 0$, so that $\bar{m}'(b) > 0$. Hence, again by the obvious analogy of Proposition 2, $\bar{m}$ is increasing in $[0, b]$, so that, by Proposition 7, $n - 1 = \bar{m} < \bar{M} = p$ in $[0, b[$ and n is strictly increasing in $[0, b]$. Now, let us consider the two following cases: (I) m is increasing in $[-a, 0]$; (II) m has a maximum at $\bar{x} \in]-a, 0[$. Next we prove that both (I) and (II) are impossible, and in this way we prove the proposition. In fact, (I) implies that M is increasing in $[-a, 0]$ also; hence $p'(x) \geqq 0$ and $n'(x) \geqq 0$ in $[-a, 0]$, so that $p(0) - 1 > \beta$ and $n(0) > \beta$. Then $\bar{M}'(0) \geqq 0$ and $\bar{M}(0) > \beta + 1$. On the other hand, $\bar{M}(b) = \beta$, so that $\bar{M}$ has a maximum in $[0, b[$, which contradicts Proposition 2. Let us consider case (II) now. If m has a maximum at $\bar{x} \in]-a, 0[$, then M is increasing in $[-a, 0]$ and m is decreasing in $[\bar{x}, 0]$, with $m'(x) < 0$ and $M'(x) > 0$ in $]\bar{x}, 0]$ (see Remark 1). Moreover, $m = p - 1$ in $[\bar{x}, 0]$, since otherwise $\bar{M}$ would have a maximum. Summing up, we can say that: (1) $n' > 0$ and $\beta < n < 1 + \beta$ in $]\bar{x}, b[$; (2) $p' < 0$ in $]\bar{x}, 0]$ and, due to Proposition 7, in $[0, b[$ $p(x) = \bar{M}(x)$ has at most one minimum where $p < \beta$, and $p' < 0$ where $p > \beta$. From (1) and (2) it follows that the equation $p(x) - n(x) = 0$ has a unique root $x^* \in]\bar{x}, b[$ and $p(x^*) = n(x^*) > \beta$. Let us now consider the function pn. Since p and n are increasing in $[-a, \bar{x}]$, we have $(pn)(\bar{x}) > (pn)(-a) = \beta(1 + \beta)$ and also $(pn)'(\bar{x}) \geqq 0$. On the other hand, multiplying (2.20) by n, (2.21) by p, and adding we obtain

$$\lambda(pn)' = C_p n + C_n p, \tag{3.10}$$

and therefore

$$\lambda(pn)'' = C_p n' + C_n p'. \tag{3.11}$$

Now, by definition of x^* we have $n \geqq p$ in $[x^*, b]$, which together with (3.8) and (3.10) yields $(pn)' > 0$ in $[x^*, b]$. Yet, from (3.11) and $C_p > 0$, $C_n < 0$, $n' > 0$ and $p' < 0$, we get $(pn)'' > 0$ in $[\bar{x}, x^*[$. Therefore pn is strictly increasing in $[\bar{x}, b]$, so that $(1 + \beta)\beta = (pn)(b) > (pn)(\bar{x}) > \beta(1 + \beta)$. This concludes the proof. □

The proof of the following result is entirely analogous.

PROPOSITION 9. *The function $\bar{m}$ is increasing in* $[0, b]$.

Finally we can prove that ψ is monotone. More precisely we have the following proposition.

PROPOSITION 10. *The function ψ is such that $\psi' > 0$ in* $[-a, b]$.

Proof. From Propositions 8 and 5 it follows that $p' \leqq 0$ in $[-a, 0]$. Hence from (2.20) we get $p\psi' = -\lambda p' + C_p > 0$, and therefore $\psi' > 0$ in $[-a, 0]$, The proof that $\psi' > 0$ in $[0, b]$ is completely analogous. □

Summary. Summing up the results of this section, we can say that the solutions (ψ, p, n) of Problem 1 have the following qualitative behavior. In $[-a, 0]$ p is decreasing, n has at most one minimum, and ψ is convex. In $[0, b]$ n is increasing, p has at most one minimum, and ψ is concave. In $[-a, b]$ p and n are positive functions bounded above by $1+\beta$ and ψ is an increasing function.

4. A convergence theorem. In this section we study the behavior of the solutions of Problem 1 as λ tends to zero. It is convenient to write explicitly the dependence of such solutions on λ, and therefore from now on we denote by $(\psi_\lambda, p_\lambda, n_\lambda)$ a solution of the problem for a given λ. Our study is based on the following a priori estimate that follows at once from the results of the previous section.

THEOREM 2. *There exists a constant C independent of λ such that for every $\lambda > 0$*

$$\|\psi_\lambda\|_{W^{2,\infty}} + \|p_\lambda\|_{W^{1,1}} + \|n_\lambda\|_{W^{1,1}} \leqq C. \tag{4.1}$$

Proof. In the following C denotes several constants, all independent of λ. Since p_λ is decreasing in $[-a, 0]$ and has no maxima in $]0, b[$ it follows that $0 < p_\lambda \leqq 1+\beta$. Analogously, we obtain $0 < n_\lambda \leqq 1+\beta$. Hence $\|p_\lambda\|_{L^1} \leqq C$ and $\|n_\lambda\|_{L^1} \leqq C$. However, since p_λ and n_λ have at most one minimum in $]-a, b[$, we obtain $\|p'_\lambda\|_{L^1} \leqq C$ and $\|n'_\lambda\|_{L^1} \leqq C$. Therefore, $\|p_\lambda\|_{W^{1,1}} \leqq C$ and $|n_\lambda\|_{W^{1,1}} \leqq C$.

From these estimates and (2.1) it follows that $\|\psi''_\lambda\|_{L^\infty} \leqq C$. From the monotonicity of ψ_λ it follows that $\|\psi_\lambda\|_{L^\infty} \leqq C$. Finally, we obtain $\|\psi'_\lambda\|_{L^\infty} \leqq C$, and therefore $\|\psi_\lambda\|_{W^{2,\infty}} \leqq C$. □

Before turning to the convergence thereom, which is in a certain sense the central result of the paper, we prove the following lemma.

LEMMA 1. *There exists a constant C independent of λ such that*

$$0 < C_p(\lambda) \leqq C\beta, \tag{4.2}$$

$$-C\beta \leqq C_n(\lambda) < 0, \tag{4.3}$$

where $C_p(\lambda)$ and $C_n(\lambda)$ are given by (2.20) *and* (2.21), *respectively. In particular,*

$$\lim_{\lambda \to 0} C_p(\lambda) = \lim_{\lambda \to 0} C_n(\lambda) = 0. \tag{4.4}$$

Proof. Let us prove (4.2), since the proof of (4.3) is similar. From (2.7) it follows that

$$e^{\alpha/\lambda} = e^{\bar{\psi}/\lambda}\sqrt{\beta/(1+\beta)}, \tag{4.5}$$

$$e^{-\alpha/\lambda} = e^{-\bar{\psi}/\lambda}\sqrt{(\beta+1)/\beta}, \tag{4.6}$$

and from this and (2.20) we get

$$C_p = (\lambda\beta\, e^{\bar{\psi}/\lambda} - \lambda(1+\beta)\, e^{-\bar{\psi}/\lambda}) \Big/ \int_{-a}^{b} e^{\psi/\lambda}. \tag{4.7}$$

Now, introducing

$$\phi := e^{\psi/\lambda}, \tag{4.8}$$

we can write

$$C_p = \left(\lambda\beta \int_{-a}^{b} \phi' - \lambda\phi(-a)\right) \Big/ \int_{-a}^{b} \phi \leqq \lambda\beta \int_{-a}^{b} \phi' \Big/ \int_{-a}^{b} \phi, \tag{4.9}$$

and since

$$\phi' = \frac{\psi'}{\lambda}\phi \tag{4.10}$$

we have

$$\int_{-a}^{b} \phi' = \frac{\psi'(\theta)}{\lambda} \int_{-a}^{b} \phi \tag{4.11}$$

with $\theta \in]-a, b[$. From (4.11), (4.9) and (4.1), (4.2) follows at once. □.

THEOREM 3. *Every sequence* $\{\psi_\lambda, p_\lambda, n_\lambda\}_{\lambda>0}$ *of solutions of Problem* 1 *converges, as* $\lambda \to 0$, *to the solution* $\{\psi_0, p_0, n_0\}$ *of the following double obstacle problem.*

PROBLEM 2. Find $\psi_0 \in C^1([-a, b])$ and ξ_p, ξ_n such that $-a \leqq -\xi_p < \xi_n \leqq b$,

$$\psi_0 = -\left(\alpha + \frac{A}{2}\right) \quad \text{in } [-a, -\xi_p], \tag{4.12}$$

$$\psi_0'' = -\operatorname{sign} x \quad \text{in }]-\xi_p, \xi_n[, \tag{4.13}$$

$$\psi_0 = \alpha + \frac{A}{2} \quad \text{in } [\xi_n, b], \tag{4.14}$$

and define

$$p_0 = 1 \quad \text{in }]-a, -\xi_p[, \qquad p_0 = 0 \quad \text{in }]-\xi_p, b[, \tag{4.15}$$

$$n_0 = 0 \quad \text{in }]-a, \xi_n[, \qquad n_0 = 1 \quad \text{in }]\xi_n, b[. \tag{4.16}$$

Proof. By Theorem 2 there exists a subsequence of $\{\psi_\lambda, p_\lambda, n_\lambda\}$—which we denote in the same way—such that

$$\psi_\lambda \to \psi_0 \quad \text{in } H^2(-a, b) \text{ weak}, \tag{4.17}$$

$$p_\lambda \to p_0 \quad \text{in } L^2(-a, b), \tag{4.18}$$

$$n_\lambda \to n_0 \quad \text{in } L^2(-a, b), \tag{4.19}$$

when $\lambda \to 0$. To prove Theorem 3 we show that these limit functions verify (4.12)-(4.16). Taking the limit as $\lambda \to 0$ in (2.1), (2.4), we can write, thanks to (4.17)-(4.19) and taking into account (2.7) and (2.8),

$$\psi_0'' = -\operatorname{sign} x - p_0 + n_0 \quad \text{in }]-a, b[, \tag{4.20}$$

$$\psi_0(-a) = -\left(\alpha + \frac{A}{2}\right), \qquad \psi_0(b) = \alpha + \frac{A}{2}. \tag{4.21}$$

However, again by Theorem 2, we have

$$\lambda p_\lambda' \to 0 \quad \text{in } L^1(-a, b), \tag{4.22}$$

$$\lambda n_\lambda' \to 0 \quad \text{in } L^1(-a, b), \tag{4.23}$$

when $\lambda \to 0$. Now taking the limit as $\lambda \to 0$ in (2.20)-(2.21), we deduce from (4.17)-(4.19) and (4.22)-(4.23) and using Lemma 1, that

$$p_0 \psi_0' = 0 \quad \text{in }]-a, b[, \tag{4.24}$$

$$n_0 \psi_0' = 0 \quad \text{in }]-a, b[. \tag{4.25}$$

Moreover, from the results of § 3 we get that

$$\psi_0'(x) \geqq 0 \quad \text{in } [-a, b], \tag{4.26}$$

$$\psi_0''(x) \geqq 0 \quad \text{in }]-a, 0[, \qquad \psi_0''(x) \leqq 0 \quad \text{in }]0, b[, \tag{4.27}$$

and we have also that

$$0 \leqq p_0 \leqq 1 \quad \text{in }]-a, b[, \tag{4.28}$$

$$0 \leqq n_0 \leqq 1 \quad \text{in }]-a, b[. \tag{4.29}$$

Now let us set

$$-\xi_p = \inf\{x \mid -a \leqq x \leqq b,\ \psi_0'(x) > 0\}, \tag{4.30}$$

$$\xi_n = \sup\{x \mid -a \leqq x \leqq b,\ \psi_0'(x) > 0\}. \tag{4.31}$$

From (4.26)-(4.27) it follows that $\xi_n, \xi_p > 0$. Hence we have (1) $\psi_0' > 0$ in $]-\xi_p, \xi_n[$ and, (2) $\psi_0' = 0$ in $]-a, -\xi_p[$ and in $]\xi_n, b[$. From (1), (4.24)-(4.25) and (4.20) it follows that ψ_0 satisfies (4.13) and $p_0 = n_0 = 0$ in $]-\xi_p, \xi_n[$, and from (2) and (4.21) it follows that ψ_0 satisfies (4.12) and (4.14). On the other hand, if $-a < -\xi_p$ we have $n_0 = 0$ and $p_0 = 1$ in $[-a, -\xi_p[$ because, from (4.20) and (4.28)-(4.29), $1 \geqq p_0 = n_0 + 1 \geqq 1$. In a similar way we prove that, if $\xi_n < b$, $p_0 = 0$ and $n_0 = 1$ in $]\xi_n, b]$. Since the limit is unique, by standard arguments we then deduce that the whole sequence $(\psi_\lambda, p_\lambda, n_\lambda)$ converges to (ψ_0, p_0, n_0). □

Remark 2. It is well known that (4.12)-(4.14) (which is clearly equivalent to (1.27)-(1.30)) can also be written as a minimum problem: Set

$$\mathbb{K} = \left\{ \phi(x) \in H^1(-a, b):\ \phi(-a) = -\left(\alpha + \frac{A}{2}\right),\right.$$
$$\left. \phi(b) = \alpha + \frac{A}{2},\ -\left(\alpha + \frac{A}{2}\right) \leqq \phi(x) \leqq \alpha + \frac{A}{2} \text{ in }]-a, b[\right\}; \tag{4.32}$$

then $\psi_0(x)$ is the unique minimum in $\mathbb{K}$ of the functional

$$J(\phi) = \frac{1}{2}\int_{-a}^{b} (\phi')^2\, dx + \int_{-a}^{b} \phi \operatorname{sign} x\, dx. \tag{4.33}$$

□

5. A uniqueness theorem. As we said in the Introduction, we are not able to prove that the solution of Problem 1 is unique for every $\lambda > 0$. In this section we prove that the solution is unique provided that λ is small enough. The proof is similar to that contained in [3] but is somewhat more involved.

THEOREM 4. *There exists λ^* such that the solution of Problem* 1 *is unique provided $\lambda < \lambda^*$.*

Proof. The scheme of the proof is the following one (as usual C denotes different constants independent of λ). First we remark that from Theorems 2 and 3 it follows that there exists a $\bar{\lambda}$ such that every solution $(\psi_\lambda, p_\lambda, n_\lambda)$ of Problem 1 satisfies

$$\|\psi_\lambda - \psi_0\|_{L^\infty} \leqq \alpha \tag{5.1}$$

whenever $\lambda < \bar{\lambda}$. Next we suppose that there exist two solutions $(\psi_{1\lambda}, p_{1\lambda}, n_{1\lambda})$, $(\psi_{2\lambda}, p_{2\lambda}, n_{2\lambda})$ and set

$$\delta_\lambda := \psi_{2\lambda} - \psi_{1\lambda}. \tag{5.2}$$

Then from the previous remark we can say that for $\lambda < \bar{\lambda}$

$$\|\delta_\lambda\|_{L^\infty} \leqq 2\alpha. \tag{5.3}$$

From this estimate we will obtain the following:

$$\|\delta_\lambda'\|_{L^2}^2 \leqq \frac{C\beta}{\lambda} \|\delta_\lambda\|_{L^\infty}. \tag{5.4}$$

From (5.4), since

$$\|\delta_\lambda\|_{L^\infty} \leqq C\|\delta'_\lambda\|_{L^2} \tag{5.5}$$

(remark that $\delta_\lambda(-a) = \delta_\lambda(b) = 0$), we get that

$$\|\delta'_\lambda\|_{L^2} \leqq \frac{C\beta}{\lambda} \tag{5.6}$$

and

$$\|\delta_\lambda\|_{L^\infty} \leqq \frac{C\beta}{\lambda}. \tag{5.7}$$

This last estimate is an improvement on (5.3). Actually from (1.23) we have $\beta = 0(\lambda^3)$, so that (5.7) can be written as

$$\|\delta_\lambda\|_{L^\infty} \leqq C\lambda^2. \tag{5.8}$$

From (5.8) we will obtain the estimate

$$\|\delta'_\lambda\|^2_{L^2} \leqq C^{\beta}_{\bar{\lambda}^2}\|\delta_\lambda\|^2_{L^\infty} \quad \text{for } \lambda \leqq \hat{\lambda} \leqq \bar{\lambda}, \tag{5.9}$$

which will enable us to conclude the proof. Indeed from (5.5) and (1.23) we can write, using (5.9),

$$\|\delta_\lambda\|^2_{L^\infty} \leqq C\lambda\, e^{-A/\lambda}\|\delta_\lambda\|^2_{L^\infty}, \tag{5.10}$$

or

$$\|\delta_\lambda\|^2_{L^\infty}(1 - C\lambda\, e^{-A/\lambda}) \leqq 0. \tag{5.11}$$

Now, since there exists $\tilde{\lambda}$ such that $1 - C\lambda\, e^{-A/\lambda} > 0$ whenever $\lambda \leqq \tilde{\lambda}$, we get $\|\delta_\lambda\|_{L^\infty} = 0$, i.e., $\psi_{1\lambda} = \psi_{2\lambda}$ for $\lambda < \lambda^* = \min\{\hat{\lambda}, \tilde{\lambda}\}$. From (2.18), (2.19) we get at once $p_{1\lambda} = p_{2\lambda}$, $n_{1\lambda} = n_{2\lambda}$.

Now it remains to prove first that (5.3) implies (5.4) and then to prove that (5.8) implies (5.9).

Proof of (5.4.). In the following we delete the subscript λ to simplify the notation. Let (ψ_1, p_1, n_1) and (ψ_2, p_2, n_2) be two solutions and let $\delta = \psi_2 - \psi_1$. From (2.10), (2.11) we write $p_i = \rho_i \exp[(\bar{\psi} - \psi_i)/\lambda]$, $n_i = \sigma_i \exp[(\bar{\psi} + \psi_i)/\lambda]$ for $i = 1, 2$, so that from the equation

$$\delta'' = p_1 - p_2 + n_2 - n_1 \tag{5.12}$$

we get

$$\begin{aligned} -\delta'' + \rho_1 \exp[(\bar{\psi} - \psi_1)/\lambda](1 - \exp(-\delta/\lambda)) - \sigma_1 \exp[(\bar{\psi} + \psi_1)/\lambda](1 - \exp(\delta/\lambda)) \\ = \exp[(\bar{\psi} - \psi_2)/\lambda](\rho_2 - \rho_1) + \exp[(\bar{\psi} + \psi_2)/\lambda](\sigma_1 - \sigma_2). \end{aligned} \tag{5.13}$$

From (2.18) and (2.19) we can write ρ_i and σ_i for $i = 1, 2$ in terms of ψ_i as follows:

$$\rho_i = \beta \exp(-2\alpha/\lambda)(1 - r_i) + \beta r_i \quad \text{with } r_i = \frac{\int_{-a}^{x} \exp(\psi_i/\lambda)}{\int_{-a}^{b} \exp(\psi_i/\lambda)}, \tag{5.14}$$

$$\sigma_i = \beta \exp(-2\alpha/\lambda)(1 - s_i) + \beta s_i \quad \text{with } s_i = \frac{\int_{x}^{b} \exp(-\psi_i/\lambda)}{\int_{-a}^{b} \exp(-\psi_i/\lambda)}. \tag{5.15}$$

Then we insert (5.14) and (5.15) in (5.13), multiply by δ, and integrate, to obtain

$$
\begin{aligned}
(5.16)\qquad \|\delta'\|_{L^2}^2 &+ \int_{-a}^{b} \delta \exp[(\bar{\psi}-\psi_1)/\lambda]\rho_1(1-\exp(-\delta/\lambda)) \\
&- \int_{-a}^{b} \delta \exp[(\bar{\psi}+\psi_1)/\lambda]\sigma_1(1-\exp(\delta/\lambda)) \\
&= \int_{-a}^{b} \delta \exp[(\bar{\psi}-\psi_2)/\lambda]\beta \exp(-2\alpha/\lambda)(r_1-r_2) \\
&+ \int_{-a}^{b} \delta \exp[(\bar{\psi}-\psi_2)/\lambda]\beta(r_2-r_1) \\
&+ \int_{-a}^{b} \delta \exp[(\bar{\psi}+\psi_2)/\lambda]\beta \exp(-2\alpha/\lambda)(s_2-s_1) \\
&+ \int_{-a}^{b} \delta \exp[(\bar{\psi}+\psi_2)/\lambda]\beta(s_1-s_2) = \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}.
\end{aligned}
$$

Let us estimate the term I; for this we write

$$
\mathrm{I} = \beta e^{-2\alpha/\lambda}\left[\frac{\exp(\bar{\psi}/\lambda)}{\int_{-a}^{b}\exp(\psi_1/\lambda)}\int_{-a}^{b}\left(\delta \exp(-\delta/\lambda)\exp(-\psi_1/\lambda)\int_{-a}^{x}\exp(\psi_1/\lambda)\right)\right.
$$
$$
\left.-\frac{\exp(\bar{\psi}/\lambda)}{\int_{-a}^{b}\exp(\psi_2/\lambda)}\int_{-a}^{b}(\exp(-\psi_2/\lambda)\int_{-a}^{x}\exp(\psi_2/\lambda))\right].
$$

Since ψ_1 and ψ_2 are increasing in $[-a, b]$, it follows that

$$
(5.17)\qquad \exp[-\psi_i(x)/\lambda]\int_{-a}^{x}\exp(\psi_i(t)/\lambda)\,dt \leqq x+a.
$$

Hence (5.3) and (5.17) yield

$$
(5.18)\qquad \mathrm{I} \leqq \beta\|\delta\|_{L^\infty}\frac{(b+a)^2}{2}\exp(\bar{\psi}/\lambda)\left[\frac{1}{\int_{-a}^{b}\exp(\psi_1/\lambda)}+\frac{\exp(-2\alpha/\lambda)}{\int_{-a}^{b}\exp(\psi_2/\lambda)}\right].
$$

Let us now estimate $\int_{-a}^{b}\exp(\psi_i/\lambda)$. Working as in the proof of Lemma 1 (see (4.8) and (4.11)), we have, for suitable $\theta_i \in]-a, b[$,

$$
(5.19)\qquad \frac{1}{\int_{-a}^{b}\exp(\psi_i/\lambda)} = \frac{\psi_i'(\theta_i)\exp(-\bar{\psi}/\lambda)}{\lambda(1-\exp(-2\bar{\psi}/\lambda))} \leqq \frac{C}{\lambda}\exp(-\bar{\psi}/\lambda)
$$

so that the following bound holds:

$$
(5.20)\qquad \mathrm{I} \leqq C\frac{\beta}{\lambda}\|\delta\|_{L^\infty}.
$$

For the next term we obtain

$$
\begin{aligned}
\mathrm{II} = \beta \exp(\bar{\psi}/\lambda)&\int_{-a}^{b}\delta[\exp(-\psi_2/\lambda)r_2-\exp(-\psi_1/\lambda)r_1] \\
&+\beta\exp(\bar{\psi}/\lambda)\int_{-a}^{b}\delta(1-\exp(-\delta/\lambda))\exp(-\psi_1/\lambda)r_1
\end{aligned}
$$

and, using (5.14), (5.17), and (5.19), we have

$$\text{II} \leqq C\frac{\beta}{\lambda}\|\delta\|_{L^\infty}+\beta\int_{-a}^{b}\delta(1-\exp(-\delta/\lambda))\exp((\bar{\psi}-\psi_1)/\lambda)r_1. \tag{5.21}$$

Analogously, we have

$$\exp[\psi_i(x)/\lambda]\int_{x}^{b}\exp(-\psi_i(t)/\lambda)\,dt \leqq b-x \tag{5.22}$$

and

$$\frac{1}{\int_{-a}^{b}\exp(-\psi_i/\lambda)} \leqq \frac{C}{\lambda}\exp(-\bar{\psi}/\lambda), \tag{5.23}$$

so that

$$\text{III} \leqq C\frac{\beta}{\lambda}\|\delta\|_{L^\infty}, \tag{5.24}$$

$$\text{IV} \leqq C\frac{\beta}{\lambda}\|\delta\|_{L^\infty}+\beta\int_{-a}^{b}\delta\exp((\bar{\psi}+\psi_1)/\lambda)(\exp(\delta/\lambda)-1)s_1. \tag{5.25}$$

Summarizing (5.16), (5.20), (5.21), (5.24), and (5.25), and recalling (5.14) and (5.15), we have

$$\begin{aligned}\|\delta'\|_{L^2}^2&+\int_{-a}^{b}\delta\exp((\bar{\psi}-\psi_1)/\lambda)\beta\exp(-2\alpha/\lambda)(1-r_1)(1-\exp(-\delta/\lambda))\\&+\int_{-a}^{b}\delta\exp((\bar{\psi}+\psi_1)/\lambda)\beta\exp(-2\alpha/\lambda)(1-s_1)(\exp(\delta/\lambda)-1)\\&\leqq C\frac{\beta}{\lambda}\|\delta\|_{L^\infty}.\end{aligned}$$

Hence (5.4) follows, since $\delta(\exp(\delta/\lambda)-1)\geqq 0$, $\delta(1-\exp(-\delta/\lambda))\geqq 0$, $1-r_1\geqq 0$, $1-s_1\geqq 0$. □

Proof of (5.9). From (2.20) and (2.21) we obtain

$$\lambda(p_2-p_1)'+\psi_2'(p_2-p_1)=c_1-p_1\delta', \tag{5.26}$$

$$\lambda(n_2-n_1)'-\psi_2'(n_2-n_1)=c_2+n_1\delta'. \tag{5.27}$$

Solving for p_2-p_1 in (5.26) and for n_2-n_1 in (5.27), we obtain

$$p_2-p_1=\frac{\exp(-\psi_2/\lambda)}{\lambda}\left(r_2\int_{-a}^{b}\exp(\psi_2/\lambda)p_1\delta'-\int_{-a}^{x}\exp(\psi_2/\lambda)p_1\delta'\right), \tag{5.28}$$

$$n_2-n_1=\frac{\exp(\psi_2/\lambda)}{\lambda}\left(s_2\int_{-a}^{b}\exp(-\psi_2/\lambda)n_1\delta'-\int_{x}^{b}\exp(-\psi_2/\lambda)n_1\delta'\right). \tag{5.29}$$

After integrating by parts in (5.28) and (5.29), we substitute them in (5.12), multiply by δ, and integrate, to obtain

$$\begin{aligned}\|\delta'\|_{L^2}^2=&\int_{-a}^{b}\left(\frac{\delta}{\lambda}\exp((\bar{\psi}-\psi_2)/\lambda)\left[-r_2\int_{-a}^{b}\delta(\exp(\delta/\lambda)\rho_1)'+\int_{-a}^{x}\delta(\exp(\delta/\lambda)\rho_1)'\right]\right)\\&+\int_{-a}^{b}\left(\frac{\delta}{\lambda}\exp((\bar{\psi}+\psi_2)/\lambda)\right.\end{aligned}$$

$$\times\left[s_2\int_{-a}^{b}\delta(\exp(-\delta/\lambda)\sigma_1)'-\int_{x}^{b}\delta(\exp(-\delta/\lambda)\sigma_1)'\right]\Big) \tag{5.30}$$
$$-\frac{1}{\lambda}\int_{-a}^{b}\delta^2(p_1+n_1)$$
$$=\mathrm{V}+\mathrm{VI}+\mathrm{VII}+\mathrm{VIII}-\frac{1}{\lambda}\int_{-a}^{b}\delta^2(p_1+n_1).$$

Then from the definition of ρ_1 ((5.14)), we have

$$\mathrm{V}=-\frac{\exp(\bar{\psi}/\lambda)}{\lambda}\int_{-a}^{b}\delta\exp(-\psi_2/\lambda)r_2\int_{-a}^{b}\delta\beta(1-\exp(-2\alpha/\lambda))\frac{\exp(\psi_2/\lambda)}{\int_{-a}^{b}\exp(\psi_1/\lambda)}$$
$$-\frac{\exp(\bar{\psi}/\lambda)}{\lambda^2}\int_{-a}^{b}\delta\exp(-\psi_2/\lambda)r_2\int_{-a}^{b}\delta\delta'\exp(\delta/\lambda)\rho_1.$$

Since $0\leqq\rho_1\leqq\beta$, from (2.14), and since (5.17), (5.19) hold, we have

$$\mathrm{V}\leqq C\frac{\beta}{\lambda^2}\|\delta\|_{L^\infty}^2+C\frac{\beta}{\lambda^3}\|\delta\|_{L^\infty}^2\int_{-a}^{b}|\delta'|\exp(\delta/\lambda) \tag{5.31}$$
$$\leqq C\frac{\beta}{\lambda^2}\|\delta\|_{L^\infty}^2,$$

where we also used the Cauchy-Schwarz inequality and (5.6), (5.7). Let us integrate by parts the following term:

$$\mathrm{VI}=\int_{-a}^{b}\left(\frac{\delta}{\lambda}\exp((\bar{\psi}-\psi_2)/\lambda)\left[\int_{-a}^{x}\delta\exp(\delta/\lambda)\rho_1'+\int_{-a}^{x}\delta\frac{\delta'}{\lambda}\exp(\delta/\lambda)\rho_1\right]\right)$$
$$=\int_{-a}^{b}\left(\frac{\delta}{\lambda}\exp((\bar{\psi}-\psi_2)/\lambda)\left[\int_{-a}^{x}\delta\exp(\delta/\lambda)\rho_1'+\frac{\delta^2}{2\lambda}\exp(\delta/\lambda)\rho_1\right.\right.$$
$$\left.\left.-\int_{-a}^{x}\frac{\delta^2}{2\lambda}(\exp(\delta/\lambda)\rho_1)'\right]\right)$$
$$=\mathrm{VI}_1+\int_{-a}^{b}\frac{\delta^3}{2\lambda^2}p_1+\mathrm{VI}_2.$$

We estimate separately VI_1 and VI_2. For the former, recalling the definition of ρ_1 and (5.17), (5.19), we get

$$\mathrm{VI}_1\leqq\|\delta\|_{L^\infty}^2\frac{\beta}{\lambda}(1-\exp(-2\alpha/\lambda))\frac{\exp(\bar{\psi}/\lambda)}{\int_{-a}^{b}\exp(\psi_1/\lambda)}\int_{-a}^{b}\left(\exp(-\psi_2/\lambda)\int_{-a}^{x}\exp(\psi_2/\lambda)\right)$$
$$\leqq C\frac{\beta}{\lambda^2}\|\delta\|_{L^\infty}^2.$$

Analogously, considering also (5.6) and (5.7), we have

$$\mathrm{VI}_2\leqq\frac{\beta}{2\lambda^2}(1-\exp(-2\alpha/\lambda))\|\delta\|_{L^\infty}^3\frac{\exp(\bar{\psi}/\lambda)}{\int_{-a}^{b}\exp(\psi_1/\lambda)}\int_{-a}^{b}\left(\exp(-\psi_2/\lambda)\int_{-a}^{x}\exp(\psi_2/\lambda)\right)$$
$$+\frac{1}{2\lambda^3}\|\delta\|_{L^\infty}^3\int_{-a}^{b}\left(\exp((\bar{\psi}-\psi_2)/\lambda)\int_{-a}^{x}|\delta'|\exp(\delta/\lambda)\rho_1\right)$$
$$\leqq C\frac{\beta}{\lambda^3}\|\delta\|_{L^\infty}^3+\frac{1}{2\lambda^3}\|\delta\|_{L^\infty}^3\int_{-a}^{b}\int_{-a}^{x}|\delta'|\exp((\psi_2(t)-\psi_2(x))/\lambda)p_1\,dt\,dx.$$

Finally, since $0 \leqq p_1 \leqq 1+\beta \leqq 2$ and ψ_2 is increasing, we get the following estimate for the term VI, using (5.6), (5.7) once more:

$$\text{VI} \leqq C\frac{\beta}{\lambda^2}\|\delta\|_{L^\infty}^2 + \int_{-a}^{b} \frac{\delta^3}{2\lambda^2} p_1. \tag{5.32}$$

Analogously,

$$\text{VII} \leqq C\frac{\beta}{\lambda^2}\|\delta\|_{L^\infty}^2 \tag{5.33}$$

and

$$\text{VIII} \leqq C\frac{\beta}{\lambda^2}\|\delta\|_{L^\infty}^2 - \int_{-a}^{b} \frac{\delta^3}{2\lambda^2} n_1. \tag{5.34}$$

Gathering (5.30)-(5.34), we obtain

$$\|\delta'\|_{L^2}^2 \leqq C\frac{\beta}{\lambda^2}\|\delta\|_{L^\infty}^2 - \frac{1}{\lambda^2}\int_{-a}^{b} \delta^2 p_1\left(\lambda - \frac{\delta}{2}\right) - \frac{1}{\lambda^2}\int_{-a}^{b} \delta^2 n_1\left(\lambda + \frac{\delta}{2}\right).$$

Since (5.8) holds there exists $\hat{\lambda} \leqq \bar{\lambda}$ such that for $\lambda \leqq \hat{\lambda}$, the two last integrals are positive; hence we obtain (5.9). □

REFERENCES

[1] F. BREZZI, *Theoretical and numerical problems in reverse biased semiconductor devices*, in Proc. of the Seventh International Symposium on Computing Methods in Applied Sciences and Engineering, Versailles, 1985.

[2] F. BREZZI, A. CAPELO, AND L. D. MARINI, *Singular perturbation problems in semiconductor devices*, in Proc. II MAS Workshop on Numer. Anal., J. P. Hennart, ed., Lecture Notes in Math., 1230, Springer-Verlag, Berlin, New York, 1986, pp. 191-198.

[3] F. BREZZI AND L. GASTALDI, *Mathematical properties of one-dimensional semiconductors*, Mat. Apl. Comp., 5 (1986), pp. 123-137.

[4] L. CAFFARELLI AND A. FRIEDMAN, *A singular perturbation problem for semiconductors*, Boll. Un. Mat. Ital. B, 7 (1987), pp. 409-421.

[5] J. W. JEROME, *Consistency of semiconductor modeling: an existence/stability analysis for the stationary van Roosbroeck system*, SIAM J. Appl. Math., 45 (1985), pp. 565-590.

[6] M. KURATA, *Numerical Analysis for Semiconductor Devices*, Lexington Press, Lexington, 1982.

[7] P. A. MARKOWICH, *The Stationary Semiconductor Device Equations*, Springer-Verlag, Vienna, 1986.

[8] P. A. MARKOWICH AND P. SZMOLYAN, *Initial transient of solution of the basic semiconductor device equations*, to appear.

[9] M. S. MOCK, *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dublin, 1983.

[10] W. V. VAN ROOSBROECK, *Theory of flow of electrons and holes in germanium and other semiconductors*, Bell. Systems J., 29 (1950), pp. 560-607.

[11] R. E. BANK, J. W. JEROME, AND D. J. ROSE, *Analytical and numerical aspects of semiconductor device modeling*, in Proc. Fifth International Conference on Computing Methods in Applied Science and Engineering, R. Glowinski and J. L. Lions, eds., North-Holland, Amsterdam, 1982.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 388-395, March 1989

007

PERSISTENCE IN INFINITE-DIMENSIONAL SYSTEMS*

JACK K. HALE† AND PAUL WALTMAN‡

Abstract. The concept of persistence reflects the survival of all components of a model ecosystem. Most of the results to date are restricted to ordinary differential equations or to dynamics on locally compact spaces. The concept is investigated here in the setting of a C^0-semigroup which is asymptotically smooth. Since the equations of population dynamics often involve delays or diffusion this seems the appropriate setting. Conditions are placed on the flow on the boundary which, given the presence of a global attractor provided by the assumption of dissipativeness and asymptotic smoothness, are necessary and sufficient for persistence.

Key words. persistence, asymptotically smooth, global attractor

AMS(MOS) subject classifications. 34C35, 34G20

1. Introduction. The notion of persistence (defined below) captures a basic idea in ecology by expressing the ultimate survival of the component populations of a model ecosystem. Most of the applications have been to autonomous systems of ordinary differential equations and thus make use of dynamics in Euclidean spaces. (Exceptions are: Dunbar, Rybakowski, and Schmitt [5] and Hutson and Moran [11], who apply this idea to reaction diffusion equations; Gatica and So [6], who work with non-autonomous ordinary differential equations; and Burton and Hutson [2], who work with delay equations.) The notion of persistence is basically a dynamic one and has been explored in the context of locally compact metric spaces in [3] and [4]. However, many of the models of population dynamics naturally involve such concepts as delays or diffusion resulting in functional differential equations or partial differential equations. There is still a natural setting for the application of the ideas of dynamical systems [7], [10] but it is in spaces that are not locally compact.

The purpose of this note is to develop this idea of persistence under a reasonable set of hypotheses so that it can be applied to biological problems with delays or diffusion. The general setting is that of a semigroup and the basic hypothesis is that of being asymptotically smooth, a notion that has been helpful in developing the dynamics in infinite-dimensional systems. The work blends the ideas of [7] and [3].

The hypotheses naturally are more difficult to check in this setting. The basic one involves knowledge of the attracting set (defined below), which is difficult to obtain in an infinite-dimensional system. Nevertheless, we feel that this is the natural and proper setting for the concept of persistence and that the theorems may be useful when specific equations are studied.

2. Preliminaries. In this section we give some background material on dynamical systems and the existence of global attractors. Let X be a complete metric space (with metric d) and suppose that $T(t): X \to X$, $t \geqq 0$, is a C^0-semigroup on X; that is, $T(0) = I$, $T(t+s) = T(t)T(s)$ for $t, s \geqq 0$, and $T(t)x$ is continuous in t, x. The positive

* Received by the editors September 28, 1987; accepted for publication (in revised form) June 20, 1988.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported by U.S. Army Research Office contract DAAL-03-86-K-0074 and National Science Foundation grant DMS-8507056.

‡ Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322. The research of this author was supported by National Science Foundation grant DMS 86-01398.

orbit $\gamma^+(x)$ through x is defined as $\gamma^+(x) = \cup_{t\geq 0}\{T(t)x\}$. The ω-limit set is defined as

$$\omega(x) = \bigcap_{\tau\geqq 0} \operatorname{Cl} \bigcup_{t\geqq\tau} \{T(t)x\}.$$

This is equivalent to saying that $y\in\omega(x)$ if and only if there is a sequence $t_n\to\infty$ as $n\to\infty$ such that $T(t_n)x\to y$ as $n\to\infty$. If B is a subset of X, we define the ω-limit set of B as

$$\omega(B) = \bigcap_{\tau\geqq 0} \operatorname{Cl} \bigcup_{t\geqq\tau} T(t)B,$$

where

$$T(t)B = \bigcup_{x\in B} \{T(t)x\}.$$

This is equivalent to saying that $y\in\omega(B)$ if and only if there are sequences $t_n\to\infty$, $x_n\in B$ such that $T(t_n)x_n\to y$ as $n\to\infty$. Note that double sequences are needed to define $\omega(B)$.

It is important to make another remark about $\omega(B)$. It is tempting to consider the set

$$\bigcup_{x\in B} \omega(x)$$

as a candidate for the limiting behavior of the set B since it contains the ω-limit set of each point. This set is generally much smaller than the set $\omega(B)$. In fact, ω-limit sets of points in B could be disconnected even when $\omega(B)$ is connected. From the point of view of the qualitative behavior of the dynamics generated by the semigroup $T(t)$, it is necessary to consider the sets $\omega(B)$ defined above.

If the points x or the sets B have negative orbits, we can define the α-limit set $\alpha(x)$ of x and α-limit set $\alpha(B)$ of B in a similar manner taking into account the possibility of multiple backward orbits. When the points or sets belong to an invariant set A, we will restrict the backward orbits to those remaining in the invariant set and denote this by $\alpha_A(x)$. Sometimes it is convenient to have the alpha limit set of a specific full orbit, $\gamma(x)$ through a point x. We denote this by $\alpha_\gamma(x)$.

A set B in X is said to be *invariant* if $T(t)B = B$ for $t\geqq 0$; that is, the mapping $T(t)$ takes B onto B for each $t\geqq 0$. This implies, in particular, that there is a negative orbit through each point of an invariant set.

A nonempty invariant subset M of X is called an *isolated invariant set* if it is the maximal invariant set of a neighborhood of itself. The neighborhood is called an *isolating neighborhood.* The *stable* (*or attracting*) set of a compact invariant set A is denoted by W^s and is defined as

$$W^s(A) = \{x \mid x\in X,\ \omega(x)\neq\phi,\ \omega(x)\subset A\}.$$

The *unstable* (*or repelling*) set, W^u is defined by

$$W^u(A) = \{x \mid x\in X, \text{there exists a backward orbit } \gamma^-(x) \text{ such that } \alpha_\gamma(x)\neq\phi,\ \alpha_\gamma(x)\subset A\}.$$

The weakly stable and unstable sets are defined as follows:

$$W^s_w(A) = \{x \mid x\in X,\ \omega(x)\neq\phi,\ \omega(x)\cap A\neq\phi\}$$
$$W^u_w(A) = \{x \mid x\in X,\ \alpha(x)\neq\phi,\ \alpha(x)\cap A\neq\phi\}.$$

A set A in X is said to be a *global attractor* if it is compact, invariant and, for any bounded set B in X, $\delta(T(t)B, A)\to 0$ as $t\to\infty$, where $\delta(B, A)$ is the distance from the set B to the set A:

$$\delta(B, A) = \sup_{y\in B} \inf_{x\in A} d(y, x).$$

In particular, this implies $\omega(B)$ exists and belongs to A. A global attractor is always a maximal compact invariant set. The semigroup $T(t)$ is said to be *asymptotically smooth* [9] if for any bounded subset B of X, for which $T(t)B \subset B$ for $t \geqq 0$, there exists a compact set K such that $\delta(T(t)B, K) \to 0$ as $t \to \infty$. In particular, $\omega(B) \subset K$. The semigroup $T(t)$ is said to be *point dissipative in* X if there is a bounded nonempty set B in X such that, for any $x \in X$, there is a $t_0 = t_0(x, B)$ such that $T(t)x \in B$ for $t \geqq t_0$.

A basic result on the existence of global attractors is the following theorem.

THEOREM 2.1 [7] [8]. *If*

(i) $T(t)$ *is asymptotically smooth,*

(ii) $T(t)$ *is point dissipative in* X,

(iii) $\gamma^+(U)$ *is bounded in* X *if* U *is bounded in* X,

then there is a nonempty global attractor A *in* X.

If there is a $t_0 \geqq 0$ such that $T(t)$ is compact for $t > t_0$ then $T(t)$ is asymptotically smooth. In this case we can dispense with hypothesis (iii) of Theorem 2.1 and prove Theorem 2.2.

THEOREM 2.2 [1]. *If*

(i) *there is a* $t_0 \geqq 0$ *such that* $T(t)$ *is compact for* $t > t_0$,

(ii) $T(t)$ *is point dissipative in* X,

then there is a nonempty global attractor A *in* X.

3. Persistence. In this section we consider a particular system motivated by biological considerations. We will assume that the metric space X is the closure of an open set X^0; that is, $X = X^0 \cup \partial X^0$, where ∂X^0 (assumed to be nonempty) is the boundary of X^0. We will also suppose that the C^0-semigroup $T(t)$ on X satisfies

$$T(t): X^0 \to X^0, \qquad T(t): \partial X^0 \to \partial X^0 \tag{3.1}$$

and let $T_0(t) = T(t)|_{X^0}$, $T_\partial(t) = T(t)|_{\partial X^0}$.

The set ∂X^0 is a complete metric space (with metric d). Therefore, if $T(t)$ satisfies the conditions of Theorem 2.1 in X, then T_∂ will satisfy the same conditions in ∂X^0. Therefore there will be a global attractor A_∂ in ∂X^0.

When $T(t)$ satisfies the conditions of Theorem 2.1 or Theorem 2.2, it may not be the case that $T_0(t)$ even has a maximal compact invariant set in X^0. There may be points x in X^0 for which $\omega(x) \cap \partial X^0 \neq \phi$. In this case, there is clearly no maximal compact invariant set in X^0. It can also happen that $\omega(x) \cap \partial X^0 = \phi$ for all x and there does not exist a maximal compact invariant set in X^0, for example, a simple predator prey model with families of periodic orbits. From the biological point of view, these concepts are important for they are related to the coexistence of populations.

The concept of persistence is introduced to assist in the understanding of these last remarks. The semigroup $T(t)$ is said to be *persistent* if for any $x \in X^0$, $\liminf_{t \to \infty} d(T(t)x, \partial X^0) > 0$. The semigroup $T(t)$ is said to be *uniformly persistent* if there is an $\eta > 0$ such that, for any $x \in X^0$, $\liminf_{t \to \infty} d(T(t)x, \partial X^0) \geqq \eta$. If $T(t)$ is persistent, it need not be uniformly persistent [4].

To obtain an equivalent formulation of persistence in terms of attractors, it is convenient to introduce two definitions. We say that a set U in X^0 is *strongly bounded in* X^0 if it is bounded in X and there is an $\eta > 0$ such that $d(x, \partial X^0) \geqq \eta$ for x in U. When we restrict the semigroup $T(t)$ to X^0 we can think of ∂X^0 as being part of "infinity." The "bounded" sets in X^0 should then be the strongly bounded sets in X. We say that $T(t)$ is *strongly point dissipative in* X^0 if there exists a strongly bounded set B in X^0 such that for any $x \in X^0$, there is a $t_0 = t_0(x, B)$ such that $T(t)x \in B$ for $t \geqq t_0$. The following proposition is obvious.

PROPOSITION 3.1. *Suppose that $T(t)$ is point dissipative in X. Then we have the following*:

(i) *$T(t)$ is persistent if and only if $\omega(x)$ is strongly bounded for each $x \in X^0$.*

(ii) *$T(t)$ is uniformly persistent if and only if $T(t)$ is strongly point dissipative in X^0.*

We need the following definition. A_0 is said to be a *global attractor for $T(t)$ in X^0 relative to strongly bounded sets* if $A_0 \subset X^0$ is compact, invariant, and $\delta(T(t)U, A_0) \to 0$ as $t \to \infty$ for all strongly bounded sets U. In particular, $\omega(U) \subset A_0$ for each strongly bounded set U in X^0. In the case in which X is a finite-dimensional Banach space (or when there is a $t_0 \geqq 0$ such that $T(t)$ is a compact for $t \geqq t_0$ and $T(t)$ is strongly point dissipative in X^0) we can use the same arguments as in the proof of Theorem 2.2 to obtain the existence of a global attractor in X^0. We can also prove Theorem 3.2.

THEOREM 3.2. *Suppose $T(t)$ satisfies* (3.1) *and we have the following*:

(i) *There is a $t_0 \geqq 0$ such that $T(t)$ is compact for $t > t_0$;*

(ii) *$T(t)$ is point dissipative in X;*

(iii) *$T(t)$ is uniformly persistent.*

Then there are global attractors A in X and A_∂ in ∂X^0 and a global attractor A_0 in X^0 relative to strongly bounded sets. Furthermore,

$$A = A_0 \cup W^u(A_\partial)$$

where

$$W^u(A_\partial) = \{x \in A \mid \alpha_A(x) \subset A_\partial\}.$$

Proof. Theorem 2.2 implies the existence of the global attractor in X and ∂X^0. With the interpretation of ∂X^0 as part of infinity for X^0 the existence of the attractor in X^0 follows from the proof of Theorem 2.2. It remains to prove that A has the specific representation stated in the theorem. Suppose $x \in A \setminus A_0$. Then there exists a full orbit $\omega(x)$ in A. Since A is compact, it follows that $\alpha_\gamma(x)$ exists, is compact, invariant, and belongs to A. Since A_0 is uniformly asymptotically stable, there is a $\delta > 0$ such that distance $d(y, A_0) \geqq \delta$ for $y \in \alpha(x)$. If $y \notin \partial X^0$, then $\omega(y) \subset A_0$. However, the previous inequality implies that this is impossible. Therefore, $\alpha(x) \subset \partial X^0$. Since $\alpha(x)$ is compact and invariant, it follows that $\alpha(x) \subset A_\partial$. The theorem is proved. □

The first hypothesis in Theorem 3.2 has little to do with the global properties of orbits. For example, it is generally satisfied for delay equations with finite delay or systems of parabolic equations. Hypothesis (ii) is a global property of the orbits, but it is expected to hold for most systems, since it essentially says that infinity is unstable. The most difficult hypothesis to verify is (iii), the uniform persistence.

We would like to verify uniform persistence by discussing only properties of the flow in a neighborhood of ∂X^0. The Lyapunov approach is an attractive way to do this, and this approach has been used in [5] and [11]. In [11] a type of Lyapunov function has been employed to show that the corresponding semigroup $T(t)$ is strongly point dissipative in X^0. Thus, a global attractor in X^0 exists from Theorem 3.2; this is a stronger conclusion that the one asserted in [11]. In the next section we give another method for obtaining uniform persistence.

If the map $T(t)$ is only asymptotically smooth then the analogue of Theorem 3.2 is Theorem 3.3.

THEOREM 3.3. *Suppose $T(t)$ satisfies* (3.1) *and we have the following*:

(i) *$T(t)$ is asymptotically smooth;*

(ii) *$T(t)$ is point dissipative in X;*

(iii) *$\gamma^+(U)$ is bounded if U is bounded in X;*

(iv) *$T(t)$ is uniformly persistent;*

(v) $\gamma^+(V)$ *is strongly bounded in* X^0 *if* V *is strongly bounded in* X^0.

Then the conclusions of Theorem 3.2 *are valid.*

4. Chains and uniform persistence. In this section we give characterizations of uniform persistence in terms of the behavior of the flow on ∂X^0. We need some additional definitions.

Let M, N be isolated invariant sets (not necessarily distinct). M is said to be *chained* to N, written $M \to N$, if there exists an element x, $x \notin M \cup N$ such that $x \in W^u(M) \cap W^s(N)$. A finite sequence $M_1, M_2, \cdots, M_k$ of isolated invariant sets will be called a *chain* if $M_1 \to M_2 \to \cdots \to M_k (M_1 \to M_1$, if $k = 1)$. The chain will be called a *cycle* if $M_k = M_1$.

The particular invariant sets of interest are

$$\tilde{A}_\partial = \bigcup_{x \in A_\partial} \omega(x). \tag{4.1}$$

$\tilde{A}_\partial$ is *isolated* if there exists a covering $M = \cup_{i=1}^k M_k$ of $\tilde{A}_\partial$ by pairwise disjoint, compact, isolated invariant sets $M_1, M_2, \cdots, M_k$ for T_∂ such that each M_i is also an isolated invariant set for T. M is called an *isolated covering*. $\tilde{A}_\partial$ will be called *acyclic* if there exists some isolated covering $M = \cup_{i=1}^k M_i$ of $\tilde{A}_\partial$ such that no subset of the M_i's forms a cycle. An isolated covering satisfying this condition will be called acyclic.

The principal result on persistence can now be stated.

THEOREM 4.1. *Suppose* $T(t)$ *satisfies* (3.1) *and we have the following*:

(i) *There is a* $t_0 \geqq 0$ *such that* $T(t)$ *is compact for* $t > t_0$;

(ii) $T(t)$ *is point dissipative in* X;

(iii) $\tilde{A}_\partial$ *is isolated and has an acyclic covering* M.

Then $T(t)$ *is uniformly persistent if and only if for each* $M_i \in M$

$$W^s(M_i) \cap X^0 = \phi. \tag{4.2}$$

For the case in which $T(t)$ is only asymptotically smooth, we have the following theorem.

THEOREM 4.2. *Suppose* $T(t)$ *satisfies* (3.1) *and we have the following*:

(i) $T(t)$ *is asymptotically smooth*;

(ii) $T(t)$ *is point dissipative in* X;

(iii) $\gamma^+(U)$ *is bounded in* X *if* U *is bounded in* X;

(iv) $\tilde{A}_\partial$ *is isolated and has an acyclic covering.*

Then the conclusions of Theorem 4.1 *are valid.*

For the proof we need the following lemma.

LEMMA 4.3. *Assume that* T *satisfies* (i)-(iii) *of Theorem* 4.2. *Let* γ_n^+ *be a sequence of precompact semiorbits, with* ω*-limit sets* ω_n. *Suppose that* M *is a compact, isolated invariant set with* $\omega_n \cap M = \phi$ *for* n *large. If* $p_n \in \omega_n$ *is such that* $d(p_n, M) \to 0$, *then there exist sequences* $\{q_n\}$, $\{r_n\}$, $q_n, r_n \in \omega_n$ *and elements* $q \in W^s(M) \backslash M$, $r \in W^u(M) \backslash M$ *with* $\lim_{n\to\infty} q_n = q$ *and* $\lim_{n\to\infty} r_n = r$. *The* q *and* r *can be found in an arbitrarily small neighborhood of* M.

Proof. Let us first prove that the set $\{\omega_n\}$ contains a subsequence that converges in the Hausdorff metric to an invariant set ω. The omega-limit set of a semiorbit γ^+ consists of full orbits and belongs to the global attractor A for $T(t)$. Let K be the set of nonempty compact subsets of $\Omega = \mathrm{Cl} \cup_{x \in X} \omega(x)$ with Hausdorff metric ρ. The subsets ω_n are uniformly bounded compact subsets of Ω. Since each element of K is a subset of A which is compact, there is a subsequence, which we again label ω_n such that $\rho(\omega_n, \omega) \to 0$, $\omega \in K$. Clearly ω is invariant.

Let U be an isolating neighborhood of M and V an open set such that $M \subset V \subset \mathrm{Cl}(V) \subset U$. Then $p_n \in V$ for large n. Since $p_n \in \omega_n$, which is invariant, there is a full orbit through p_n. Since $p_n \in V$, there is a corresponding $y_n \in \partial V$, with $T(\tau_n)p_n = y_n$ and $T(t)p_n \in V$, for $0 > t > \tau_n$. Since $y_n \in \Omega$ we may select a convergent subsequence, again called y_n, such that $\lim_{n\to\infty} y_n = y \in \partial V \cap \omega$ and, in particular, $y \notin M$.

If $\{\tau_n\}$ were bounded, we could select a convergent subsequence $\tau_n \to \tau$ which, by continuity, makes $T(\tau)y \in M$. Hence $T(t)y \in M$ for $t > \tau$, and $y \in W^s(M) \backslash M$. Thus we may assume that $\tau_n \to -\infty$. This has the consequence that $\gamma^+(y) \subset V$ or $y \in W^s(M)$.

Since $y \in \omega$ we may select the q_n's as stated. The proof for the r_n's follows similarly, taking into account the possibility of multiple backward orbits. (Note that in this case we need only one backward orbit to remain in V.) This completes the proof of the lemma. □

Remark 4.4. If, in Lemma 4.3, there is a single orbit $\gamma^+(x)$, with $p_n \in \gamma^+(x)$ and $\lim_{t\to\infty} d(p_n, M) = 0$, and $x \notin W^s(M)$, the proof provides similar sequences $\{r_n\}$ and $\{q_n\}$ on $\gamma^+(x)$ (see [3, Thm. 4.1] and [5, Thm. 2.2]).

Proof of Theorem 4.2. The necessity of (4.2) is clear. Suppose (4.2) holds and $T(t)$ is not uniformly persistent. There are two cases to be considered in the proof. There is a sequence of points p_n with either $p_n = \gamma^+(t_n)$ for some orbit $\gamma^+(x)$, $x \in X^0$ or $p_n \in \omega_n$ for some sequence of omega limits sets, such that $d(p_n, \partial X^0) \to 0$ as $n \to \infty$. Choose a subsequence such that $p_n \to q$ and, if in the second case, such that $\omega_n \to \omega$ as in the proof of Lemma 4.3. Let Ω be $\omega(x)$ in the first case or the ω of Lemma 4.3 in the second. Clearly, $\gamma^+(q) \subset \partial X^0$ and $q \in W^s_w(M_i)$ for some M_i, say for M_1. By Lemma 4.3 (or Remark 4.4), there exists a point $q_1 \in W^s(M_1) \cap \Omega$. Since Ω is invariant, there exists a full orbit $\gamma(q_1)$ through q_1 that lies in Ω. $\alpha_\gamma(q_1)$ exists, and, since $\alpha_\gamma(q_1)$ is invariant, $\alpha_\gamma(q_1) \cap W^u_w(M_j) \neq \phi$, say for $j = 2$. If $\alpha_\gamma(q_1) \subset M_2$, then M_2 is chained to M_1. Clearly we can choose a new sequence of points p_n either on the sets ω_n or on the orbit $\gamma^+(x)$ whose distance from M_2 tends to zero. Using Lemma 4.3 or Remark 4.4, choose q_2 and repeat the argument. If $\alpha_\gamma(p_1)$ is not a proper subset of M_2, then we proceed essentially as in the proof of [3, Thm. 3.1] to reach a contradiction of the no cycle condition. The proof now follows the general scheme of the proof of Theorem 3.1 of [3], keeping in mind the two cases mentioned above and taking care to use a full orbit when constructing the alpha limit sets needed to chain sets. The reference in the proof in [3], [4], however, is not needed in view of the statement of Lemma 4.3. □

5. An example. Suppose that $x(t)$ and $y(t)$ represent populations that grow according to the delayed logistic equation

$$x'(t) = r_1 x(t)(1 - x(t-1)), \qquad y'(t) = r_2 y(t)(1 - y(t-1)).$$

If r_1 and r_2 are sufficiently small, each population grows to the carrying capacity [12, p. 14]—in this case, to one. Suppose that each produces a metabolic product or a toxin that inhibits the growth of the other, i.e., each reduces the current intrinsic growth rate of its competitor. The equations then take the form

$$(5.1) \qquad x'(t) = r_1 x(t)[1 - x(t-1) - \mu_1 y(t)], \qquad y'(t) = r_2 y(t)[1 - y(t-1) - \mu_2 x(t)].$$

This is one of a large variety of possible competition equations with delays that may be modeled on the Lotka–Volterra equations. We show that if, in addition to the above assumption on the r_i's, μ_1 and μ_2 are less than one, the system is persistent.

The appropriate space X is the positive cone of $C[0,1] \times C[0,1]$. The form of the system of (5.1) insures that solutions corresponding to positive initial conditions (positive functions on $[-1, 0]$) remain positive, i.e., the positive cone X is invariant.

For any pair of initial functions $(\phi, \psi) \in X$, let $x(t, \phi, \psi)$ be the solution of (5.1) and define $T(t)(\phi, \psi) \in X$, $t \geqq 0$, by $T(t)(\phi, \psi)(\theta) = (x(t+\theta, \phi, \psi), y(t+\theta, \phi, \psi))$, $-1 \leqq \theta \leqq 0$.

Then $T(t)$ is completely continuous for $t > 1$. Simple comparison arguments show that $T(t)$ is point dissipative. There are three constant solutions on the boundary of X corresponding to $x(t) = 0$ and $y(t) = 0$, $x(t) = 1$ and $y(t) = 0$, and $x(t) = 0$ and $y(t) = 1$. The origin is clearly unstable. Let us linearize around $(0, 1)$ to find

$$h_1'(t) = r_1 h_1(t)(1-\mu_1), \qquad h_2'(t) = r_2[-h_2(t-1) - \mu_2 h_1(t)].$$

The characteristic equation factors as

$$[\lambda - r_1(1-\mu_1)][\lambda + r_2 e^{-\lambda}] = 0.$$

There is a unique positive (since $\mu_1 < 1$) eigenvalue corresponding to the first square bracket being set equal to zero. This corresponds to an eigenvector of the form $[1, 0]^T$. The eigenvalues from the second square bracket can be shown to have negative real parts. These correspond to an eigenvector of the form $[0, 1]^T$, and hence correspond to solutions that remain in the part of the boundary of X given by $x(t) \equiv 0$. Thus the stable set of this constant solution does not intersect the interior of the cone X. A similar argument applies at the constant solutions $x(t) = 1$, $y(t) = 0$, when $\mu_2 < 1$. Since the origin is unstable, $\tilde{A}_\partial$ is just the union of the constant solutions. Taking the M_i's to be these constant solutions, there are no cycles in the boundary of X. The constant solutions are also isolated invariant sets. Hence (5.1) is uniformly persistent, and there is a global attractor in the interior of $C[0, 1] \times C[0, 1]$.

When the number of equilibrium points is small (as expected in population equations) and the dimension of the unstable manifolds is small, we can anticipate checking the chain condition directly as in the example by letting the sets M_i be the equilibrium points. The results, however, allow for greater flexibility in the choice of the cover. For example, if one face had three equilibria P_i, $i = 1, 2, 3$, and connecting orbits γ_{ij} for $(i, j) \in I = \{(1, 2), (2, 3), (3, 1)\}$, connecting P_i to P_j, then we could select as an element of the cover the three points and the connecting orbits. We would then need to show that the attracting set of this element of the cover does not intersect the interior of the space. Often there can be a good biological reason that prevents cycles. In the example above, the instability of the origin (common for competition problems) precludes the connection of invariant sets in the two faces $(x, 0)$ and $(0, y)$, $x, y \in C[0, 1]$. In predator-prey problems there are no compact invariant sets on the predator axis (corresponding to extinction in the absence of prey), which is often helpful. There are obvious difficulties when the limit sets are more complex than equilibria or periodic orbits. Ideally we would like to devise a criterion that would make the sets $\tilde{A}_\partial$ "uniformly repelling." $\tilde{A}_\partial$, of course, if it is isolated, is a one-element cover so, in principle, the theorem can be applied directly to it.

REFERENCES

[1] J. BILLOTI AND J. P. LASALLE, *Periodic dissipative processes*, Bull. Amer. Math. Soc., 6 (1971), pp. 1082-1089.

[2] T. A. BURTON AND V. HUTSON, *Repellers in systems with infinite delay*, J. Math. Anal. Appl., to appear.

[3] G. BUTLER AND P. WALTMAN, *Persistence in dynamical systems*, J. Differential Equations, 63 (1986), pp. 255-263.

[4] G. BUTLER, H. I. FREEDMAN AND P. WALTMAN, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425-430.

[5] S. R. DUNBAR, K. F. RYBAKOWSKI, AND K. SCHMITT, *Persistence in models of predator-prey populations with diffusion*, J. Differential Equations, 65 (1986), pp. 117-138.

[6] J. A. GATICA AND J. W.-H. SO, *Predator-prey models with almost periodic coefficients*, Applicable Analysis, 27 (1988), pp. 143-152.

[7] J. K. HALE, *Asymptotic behavior and dynamics in infinite dimensions*, in Nonlinear Differential Equations, J. K. Hale and P. Martinez-Amores, eds., Pitman, Marshfield, MA, 1986.

[8] ———, *Asymptotic Behavior of Dissipative Systems*, Mathematics Surveys and Monographs, American Mathematical Society, Providence, RI, 1988.

[9] J. K. HALE, J. P. LASALLE, AND M. SLEMROD, *Theory of a general class of dissipative processes*, J. Math. Anal. Appl., 39 (1972), pp. 177-191.

[10] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, New York, 1981.

[11] V. HUTSON AND W. MORAN, *Repellers in reaction-diffusion equations*, Rocky Mountain J. Math., 17 (1987), pp. 301-314.

[12] V. B. KOLMANOVSKII AND V. R. NOSOV, *Stability of Functional Differential Equations*, Academic Press, Orlando, FL, 1986.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 396-407, March 1989

008

SOMMERFELD DIFFRACTION PROBLEMS WITH FIRST AND SECOND KIND BOUNDARY CONDITIONS*

F.-O. SPECK†

Abstract. A class of interface problems is considered where the solutions of one Helmholtz equation in the upper half-plane $x_2 > 0$ and of another in the lower half-plane $x_2 < 0$, respectively, are related by two transmission conditions on the line $x_2 = 0$. One of these is of the first kind, the other one of the second kind, and they are different for $x_1 > 0$ and $x_1 < 0$, respectively. In general, there appears a coupled system of Wiener-Hopf equations. Necessary and sufficient conditions for the correctness of the problem in a Sobolev space setting are presented as well as explicit formulas for a factorization of the Fourier symbol matrix of the one-medium problem, the solution in closed form, and its asymptotics near the origin.

Key words. diffraction problem, Wiener-Hopf operator, factorization, matrix function, Helmholtz equation

AMS(MOS) subject classifications. 47B35, 35J05, 45E10

1. Introduction. We investigate a certain class of diffraction problems $\mathcal{P}$ leading to simultaneous 2×2 systems of Wiener-Hopf equations. These can be considered in two ways, which we combine in this paper.

First, the classical Wiener-Hopf technique, represented by Noble [17], for instance, centers around a *function theoretic factorization* of the Fourier symbol matrix function $\sigma(\xi)$, $\xi \in \mathbb{R}$, i.e., an *explicit representation* $\sigma = \sigma_- \sigma_+$ with upper/lower holomorphic matrix functions $\sigma_\pm(\xi)$ for $\operatorname{Im} \xi \gtrless 0$ with algebraic behavior at infinity. In contrast to factoring rational matrices [1], this is a very subtle problem (see Heins [6], [7], Daniele [2], [3], Rawlins [19], Khrapkov [10], and others [8], [9], [13], [14], [23]), and it seems to be the only method for obtaining a *closed form solution.*

On the other hand, we may be interested in finding a function space setting such that $\mathcal{P}$ is well posed. This is equivalent to a certain *operator factorization* type in a more general context (see Devinatz and Shinbrot [4], Mikhlin and Prössdorf [16], and Speck [21]) where some invariant subspaces are involved.

For the one-medium problems we will find a class of symbol matrices σ depending on a parameter λ that can be factored by Daniele's method—provided λ is not exceptional (see Theorem 3.1). It is remarkable that this factorization does not correspond to a factorization of a bounded operator into bounded operators (Lemma 4.1); for only a small subclass of problems (a parameter subset $(0, 1) \subset \mathbb{C}$ of measure zero) can this be achieved by rearrangement (Theorem 4.4). For the majority of problems there exists only a *generalized factorization* corresponding to an operator factorization into *unbounded operators* (Corollary 4.6), which is best understood in the theory of systems of Cauchy type singular integral equations (see [16]). However, it helps to solve the well-posed problem (Corollary 5.1), to find the asymptotic behavior (Theorem 5.2) and to illuminate the various structures of $\mathcal{P}_\lambda$ in a function space setting based on the energy norm (Theorem 2.1 and Corollary 4.6).

* Received by the editors August 10, 1987; accepted for publication June 14, 1988. This work was sponsored by the Deutsche Forschungsgemeinschaft under grant Me 261/4-1.

† Fachbereich Mathematik, Technische Hochschule Darmstadt, D-6100 Darmstadt, Federal Republic of Germany.

We now formulate problem $\mathscr{P}$. Find

$$
\begin{aligned}
&u \in L^2(\mathbb{R}^2),\\
&u_{|\Omega^\pm} \in H^1(\Omega^\pm), \qquad \Omega^\pm\colon x_2 \gtrless 0,\\
&(\Delta + k_1^2)u = 0 \quad \text{in } \Omega^+,\\
&(\Delta + k_2^2)u = 0 \quad \text{in } \Omega^-
\end{aligned} \tag{1.1}
$$

where $\operatorname{Im} k_j > 0$ holds and the Dirichlet data $u_0^\pm = u_{|x_2=\pm 0}$ and the Neumann data $u_1^\pm = \partial u/\partial x_{2|_{x_2=\pm 0}}$ satisfy

$$
\begin{aligned}
&a_0 u_0^+ + b_0 u_0^- = h_0\\
&a_1 u_1^+ + b_1 u_1^- = h_1
\end{aligned} \quad \text{on } \mathbb{R}_+,
\qquad
\begin{aligned}
&a_0' u_0^+ + b_0' u_0^- = h_0'\\
&a_1' u_1^+ + b_1' u_1^- = h_1'
\end{aligned} \quad \text{on } \mathbb{R}_- \tag{1.2}
$$

with known constant coefficients $a_0, \dots, b_1' \in \mathbb{C}$ and $h_0 \in H^{1/2}(\mathbb{R}_+)$, $h_1 \in H^{-1/2}(\mathbb{R}_+)$, $h_0' \in H^{1/2}(\mathbb{R}_-)$, $h_1' \in H^{-1/2}(\mathbb{R}_-)$.

The following facts are known from [22], which is used for reference.

Let $1_\pm$ denote the characteristic function of $\mathbb{R}_\pm$, F the Fourier transformation, $\hat{u}(\xi_1, \xi_2) = Fu(x_1, x_2) = \int \exp(i(\xi_1 x_1 + \xi_2 x_2))u(x_1, x_2)d(x_1, x_2)$ and $t_j(\xi_1) = (\xi_1^2 - k_j^2)^{1/2}$, $j = 1, 2$, with branch cuts $\pm k_j \pm i\tau$, $\tau \geqq 0$, and $t_j(\xi_1) \sim \xi_1$ at $+\infty$. For brevity ξ_1 is replaced by ξ in what follows.

First, any solution of (1.1) satisfies the *representation formula*

$$
\begin{aligned}
u(x_1, x_2) &= G\begin{pmatrix} u_0^+ \\ u_0^- \end{pmatrix}(x_1, x_2)\\
&= F^{-1}_{\xi \mapsto x_1}\{e^{-x_2 t_1(\xi)}\hat{u}_0^+(\xi)1_+(x_2) + e^{x_2 t_1(\xi)}\hat{u}_0^-(\xi)1_-(x_2)\}.
\end{aligned} \tag{1.3}
$$

Second, we introduce the *boundary operators*

$$
\begin{aligned}
&B_\pm\colon H^{1/2}(\mathbb{R})^2 \to H^{1/2}(\mathbb{R}) \times H^{-1/2}(\mathbb{R}),\\
&B_+ = F^{-1}\begin{pmatrix} a_0 & b_0 \\ -a_1 t_1 & b_1 t_2 \end{pmatrix} \cdot F = F^{-1}\sigma_{B_+} \cdot F,\\
&B_- = F^{-1}\begin{pmatrix} a_0' & b_0' \\ -a_1' t_1 & b_1' t_2 \end{pmatrix} \cdot F = F^{-1}\sigma_{B_-} \cdot F,
\end{aligned} \tag{1.4}
$$

which correspond to (1.2) but act on functions living on the whole axis. Problem $\mathscr{P}$ is said to be of *normal type*, if $\sigma_{B_\pm}(\xi)$ are regular for $\xi \in \mathbb{R}$ and both are *stable at infinity*:

$$
[\det \sigma_{B_\pm}(\xi)]^{\pm 1} = O(|\xi|^{\pm 1}), \qquad |\xi| \to \infty. \tag{1.5}
$$

This assumption is equivalent to the bijectivity of $B_\pm$ and can also be written in the following form:

$$
\begin{aligned}
&a_0 b_1 t_2(\xi) + b_0 a_1 t_1(\xi) \neq 0, \qquad \xi \in \mathbb{R},\\
&a_0 b_1 + b_0 a_1 \neq 0
\end{aligned} \tag{1.6}
$$

supplemented by analogous conditions for the primed coefficients.

Third, there is the following theorem.

EQUIVALENCE THEOREM. *Let $\mathscr{P}$ be of normal type (or only B_- invertible). Then u solves Problem $\mathscr{P}$, if and only if* (i) *u is represented by* (1.3) *where $u_0^\pm$ are given by*

$$
\begin{pmatrix} u_0^+ \\ u_0^- \end{pmatrix} = B_-^{-1}\left\{\begin{pmatrix} v_+ \\ w_+ \end{pmatrix} + \begin{pmatrix} l_e h_0' \\ l_o h_1' \end{pmatrix}\right\} \tag{1.7}
$$

with even or odd extension of the data onto the whole axis and (ii) *the functions* (*more precisely*: *functionals*) v_+, w_+ *are solutions of the Wiener–Hopf system*

$$W\begin{pmatrix} v_+ \\ w_+ \end{pmatrix} = \begin{pmatrix} h_0 \\ h_1 \end{pmatrix} - 1_+ \cdot B_+ B_-^{-1} \begin{pmatrix} l_e h_0' \\ l_o h_1' \end{pmatrix} = \begin{pmatrix} h_0^* \\ h_1^* \end{pmatrix},$$

$$W = 1_+ \cdot F^{-1}\sigma \cdot F : \tilde{H}^{1/2}(\mathbb{R}_+) \times \tilde{H}^{-1/2}(\mathbb{R}_+) \to H^{1/2}(\mathbb{R}_+) \times H^{-1/2}(\mathbb{R}_+), \tag{1.8}$$

$$\sigma = \sigma_{B_+}\sigma_{B_-}^{-1}.$$

Note that $\tilde{H}^s(\mathbb{R}_+)$ is defined to be the closed subspace of $H^s = H^s(\mathbb{R})$ functions supported on $\overline{\mathbb{R}_+}$ (the natural embedding into $H^s(\mathbb{R}_+)$ is not closed with respect to the $H^s(\mathbb{R}_+)$ topology for $s = \frac{1}{2} \bmod 1$). This space is sometimes denoted by $H^s_{00}(\mathbb{R}_+)$ for $s = \mu + \frac{1}{2}$, $\mu \in \mathbb{N}_0$ [12], H_s^+ [5], $\tilde{W}^{s,2}(\mathbb{R}_+)$, or $H^+_{s,2}(\mathbb{R})$ [20]. The multiplication operator 1_+ acts on distributions as a restriction on test functions supported on $\overline{\mathbb{R}_+}$. Other (continuous) extensions instead of l_e, l_o are admissible.

2. The structure of Wiener–Hopf operators corresponding to problem $\mathscr{P}$.

THEOREM 2.1. *Let $\mathscr{P}$ be of normal type. The operator W given by* (1.8) *is Fredholm, if and only if the elements of the matrix*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \frac{1}{a_0'b_1' + b_0'a_1'} \begin{pmatrix} a_0b_1' + b_0a_1' & -a_0b_0' + b_0a_0' \\ -a_1b_1' + b_1a_1' & a_1b_0' + b_1a_0' \end{pmatrix} \tag{2.1}$$

satisfy the conditions

$$ad = 0, \qquad bc \neq 0 \tag{2.2}$$

or

$$ad \neq 0, \qquad \lambda^{-1} = bc/ad \notin [0, 1]. \tag{2.3}$$

In both cases, the analytical index of W equals zero (*we call $\lambda = ad/bc$ the characteristic parameter of $\mathscr{P}$*).

Proof. The symbol matrix σ reads explicitly

$$\sigma = \frac{1}{a_0'b_1't_2 + b_0'a_1't_1} \begin{pmatrix} a_0b_1't_2 + b_0a_1't_1 & -a_0b_0' + b_0a_0' \\ (-a_1b_1' + b_1a_1')t_1t_2 & a_1b_0't_1 + b_1a_0't_2 \end{pmatrix}. \tag{2.4}$$

Now put $t(\xi) = (\xi^2 - k^2)^{1/2}$ with an auxiliary wave number k (that we only need in the case $k_1 \neq k_2$) and

$$t(\xi) = t_-(\xi)t_+(\xi) = (\xi - k)^{1/2}(\xi + k)^{1/2} \tag{2.5}$$

with the usual choice of branches, see above. It is known [5, Thm. 4.4, Lemma 4.6], [22] that the mappings

$$\begin{aligned} &F^{-1}t_+^r \cdot F : \tilde{H}^s(\mathbb{R}_+) \to \tilde{H}^{s-r/2}(\mathbb{R}_+), \\ &1_+ \cdot F^{-1}t_-^r \cdot Fl : H^s(\mathbb{R}_+) \to H^{s-r/2}(\mathbb{R}_+) \end{aligned} \tag{2.6}$$

are bijective for any extension l and arbitrary $r, s \in \mathbb{R}$. Therefore W is equivalent to (coincides up to invertible factors with) the *lifted operator*

$$W_0 = 1_+ \cdot F^{-1}\sigma_0 \cdot F : L^2(\mathbb{R}_+)^2 \to L^2(\mathbb{R}_+)^2 \tag{2.7}$$

where the *lifted matrix*

$$\begin{aligned} \sigma_0 &= \begin{pmatrix} t_- & 0 \\ 0 & t_-^{-1} \end{pmatrix} \sigma \begin{pmatrix} t_+^{-1} & 0 \\ 0 & t_+ \end{pmatrix} \\ &= \frac{1}{a_0'b_1't_2 + b_0'a_1't_1} \begin{pmatrix} (a_0b_1't_2 + b_0a_1't_1)(t_-/t_+) & (-a_0b_0' + b_0a_0')t \\ (-a_1b_1' + b_1a_1')(t_1t_2/t) & (a_1b_0't_1 + b_1a_0't_2)(t_+/t_-) \end{pmatrix} \end{aligned} \tag{2.8}$$

is also invertible in $C(\mathbb{R})^{2\times 2}$ and tends to

$$\sigma_0(\pm\infty)=\begin{pmatrix}\pm a & b\\ c & \pm d\end{pmatrix} \tag{2.9}$$

at infinity, i.e., σ_0 is *piecewise continuous* on the (one-point-)compactified axis $\dot{\mathbb{R}}$.

The theory of systems of Cauchy-type singular integral equations on the unit circle, which can be transferred to the real line case by the stereographic projection (or Cayley transformation) [16, Chaps. V, IV.6], implies that the Fredholm property of W_0 (and thus of W) is equivalent to

$$\begin{aligned}&\det \sigma_0(\xi)\neq 0, \qquad \xi\in\mathbb{R},\\ &\det[\mu\sigma_0(-\infty)+(1-\mu)\sigma_0(+\infty)]\neq 0, \qquad \mu\in[0,1].\end{aligned} \tag{2.10}$$

This can easily be rewritten in the form (2.2)-(2.3). Furthermore, we obtain the index formula

$$\begin{aligned}\operatorname{Ind} W&=\dim N(W)-\operatorname{codim} R(W)\\ &=\operatorname{Ind} W_0=-\operatorname{ind}\det\sigma_0=0\end{aligned} \tag{2.11}$$

where the winding number of $\det\sigma_0$ vanishes, since

$$\begin{aligned}\det\sigma_0(\xi)&=\det\sigma(\xi)=\det\sigma_{B_+}(\xi)/\det\sigma_{B_-}(\xi)\\ &=(a_0b_1t_2(\xi)+b_0a_1t_1(\xi))/(a_0'b_1't_2(\xi)+b_0'a_1't_1(\xi))\end{aligned} \tag{2.12}$$

is an even function in ξ.

COROLLARY 2.2. *There exists a generalized factorization* [16, V.5-6] *in* $L^2(\mathbb{R})^2$

$$\sigma_0(\xi)=\sigma_{0-}(\xi)\begin{pmatrix}\left(\dfrac{\xi-i}{\xi+i}\right)^{\kappa_1} & 0\\ 0 & \left(\dfrac{\xi-i}{\xi+i}\right)^{\kappa_2}\end{pmatrix}\sigma_{0+}(\xi), \tag{2.13}$$

i.e., with $\kappa_1,\kappa_2\in\mathbb{Z}$, $\sigma_{0\pm}^{\pm 1}\in L^2(\mathbb{R},\rho)^{2\times 2}$ *are holomorphically extendable into the upper/lower complex half-plane where* $\rho(\xi)=(\xi^2+1)^{-1/2}$ *holds (the elements of* $\rho\sigma_{0\pm}^{\pm 1}$ *belong to* $L^2(\mathbb{R})$ *where* $\sigma_{0\pm}^{-1}$ *denote the inverse matrices), if and only if* $\mathcal{P}$ *is of normal type and* (2.2)-(2.3) *are satisfied. This also yields* $\operatorname{Ind} W=-\kappa_1-\kappa_2=0$.

Remarks 2.3. (1) There corresponds a *bounded operator factorization* $A_0=F^{-1}\sigma_0\cdot F=A_{0-}CA_{0+}$, $A_{0\pm}$, $C\in\mathcal{L}(L^2(\mathbb{R})^2)$, with (2.13), if and only if the matrix functions are bounded invertible, i.e., $\sigma_{0\pm}^{\pm 1}\in L^\infty(\mathbb{R})^{2\times 2}$. This is also called *a cross factorization* in the context of general Wiener-Hopf operators [21], which yields a generalized inverse of W_0 (cf. [16, Thm. 4.2].). If the additional factors satisfy $\sigma_{0\pm}\in C(\dot{\mathbb{R}})^{2\times 2}$, then (2.13) is said to be a *right canonical (matrix) factorization* [16, V.3.2.]. But here this is not true in general according to discontinuities at infinity (see (2.9)).

(2) So far we do not know about $\kappa_1=\kappa_2=0$ and the invertibility of W. The previous result is not constructive.

(3) Obviously, the system decomposes, i.e., σ is triangular (see (2.4)) if and only if (i) the vectors (a_0,b_0) and (a_0',b_0') or the vectors (a_1,b_1) and (a_1',b_1') are parallel, which means that the corresponding data are given along the full line, or (ii) $a_0=a_1'=0$ (or similar) according to a pure boundary value problem for the lower (or upper) half-plane, which can be treated separately, or (iii) in the case $t_1=t_2$ only: the vectors (a_0,b_0) and (a_1',b_1') or (a_1,b_1) and (a_0',b_0') are antiparallel. The first case (i) $bc=0$ is excluded by (2.2)-(2.3). The range of W is not closed in $H^{1/2}(\mathbb{R}_+)\times H^{-1/2}(\mathbb{R}_+)$;

there appear compatibility conditions and problems that are well-posed in other space settings where at most a single Wiener-Hopf equation must be solved. The other two cases do not disturb the conditions (2.2)-(2.3) and we have an invertible system according to two isolated equations of the type that appears for Sommerfeld's half-plane problem. All these cases have been discussed in [22].

3. The explicit function theoretic factorization for one-medium problems. We consider the nondecomposing case of a normal type problem $\mathcal{P}$ with $t_1 = t_2 = t$ and write σ from (2.4) in the standard form

$$\sigma = \begin{pmatrix} a & bt^{-1} \\ ct & d \end{pmatrix} \to \begin{pmatrix} 1 & t^{-1} \\ \lambda^{-1}t & 1 \end{pmatrix} \tag{3.1}$$

after some elementary transformations.

The following factorization of this nonrational matrix function is based on the work of Daniele [2] and others (see above). Only the case $\lambda = -1$ corresponding to the mixed Dirichlet-Neumann problem

$$\begin{aligned} &u_0^+ = h_0, \quad u_1^- = h_1 \quad \text{on } \mathbb{R}_+, \\ &u_0^+ - u_0^- = u_1^+ - u_1^- = 0 \quad \text{on } \mathbb{R}_- \end{aligned} \tag{3.2}$$

has been solved before [7], [14], [18]. The exceptional (decomposing) case $\lambda = 0$ or ∞ is excluded here, $\lambda = 1$ is nonnormal and $\lambda = 2$ corresponds to the strange reference problem where, for instance, u_0^+, $u_1^+ - u_1^- \in \mathbb{R}_+$ and u_1^-, $u_0^+ - u_0^-$ on $\mathbb{R}_-$ are given, (see [22]) and where W_0 is not Fredholm in our (L^p, $p = 2$) setting (see (2.3)).

THEOREM 3.1. *For any $\lambda \in \mathbb{C}$, $0 \neq \lambda \neq 1$, a factorization $\sigma = \sigma_- \sigma_+$ into lower/upper holomorphic function matrices is given by*

$$\begin{aligned} \sigma_\pm = (1-\lambda^{-1})^{1/4} \Big\{ &\cosh [C \cdot \log \gamma_\pm] \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &- \sinh [C \cdot \log \gamma_\pm] \begin{pmatrix} 0 & \sqrt{\lambda}/t \\ t/\sqrt{\lambda} & 0 \end{pmatrix} \Big\} \end{aligned} \tag{3.3}$$

with

$$C = \frac{i}{\pi} \log \frac{\sqrt{\lambda}+1}{\sqrt{\lambda}-1}, \qquad \gamma_\pm(\xi) = \frac{\sqrt{k \pm \xi} + i\sqrt{k \mp \xi}}{\sqrt{2k}} \tag{3.4}$$

and algebraic behavior as $\xi \to \pm\infty$ (e.g., put $\arg \sqrt{\lambda} \in [0, \pi)$, $\arg (\sqrt{\lambda}+1)/(\sqrt{\lambda}-1) \in [-\pi, 0]$, $\operatorname{Re} C \in [0, 1]$, $\arg \sqrt{k} = \frac{1}{2} \arg k$ *and branch cuts as before*).

Proof. To construct $\sigma_\pm$ we write [2]

$$\sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{\lambda t} \begin{pmatrix} 0 & \lambda \\ t^2 & 0 \end{pmatrix} \tag{3.5}$$

where the latter matrix has polynomial elements and find a matrix $M = \log \sigma$ in the sense of $\sigma = \exp M$ (which is defined as a series) by putting

$$\log \sigma = \frac{1}{2} \log g \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{2} \tau \begin{pmatrix} 0 & \lambda \\ t^2 & 0 \end{pmatrix} \tag{3.6}$$

with

$$g = \det \sigma = 1 - \lambda^{-1}, \qquad \tau = \frac{-1}{\sqrt{\lambda}\, t} \log \frac{\sqrt{\lambda}+1}{\sqrt{\lambda}-1}. \tag{3.7}$$

Surely, τ does not depend on the branch of $\sqrt{\lambda}$. An additive decomposition

$$\begin{aligned}\frac{1}{t(\xi)} &= \frac{-2i}{\pi t(\xi)}\left\{\log\frac{\sqrt{k+\xi}+i\sqrt{k-\xi}}{\sqrt{2k}}+\log\frac{\sqrt{k-\xi}+i\sqrt{k+\xi}}{\sqrt{2k}}\right\} \\ &= \frac{-2i}{\pi t(\xi)}\{\log\gamma_+(\xi)+\log\gamma_-(\xi)\}\end{aligned} \tag{3.8}$$

($\gamma_\pm$ are not upper/lower functions themselves) yields

$$\tau = \tau_+ + \tau_-, \qquad \tau_\pm(\xi) = \frac{2C}{\sqrt{\lambda}\, t(\xi)}\log\gamma_\pm(\xi) \tag{3.9}$$

where $\tau_\pm$ are holomorphic in $\operatorname{Im}\xi > -\operatorname{Im} k$ and $\operatorname{Im}\xi < \operatorname{Im} k$, respectively. A factorization of $g = g_- g_+$, $g_\pm = (1-\lambda^{-1})^{1/2}$ is trivial here.

Associating

$$(\log\sigma)_\pm = \frac{1}{2}\log g_\pm\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} + \frac{1}{2}\tau_\pm\begin{pmatrix}0 & \lambda\\ t^2 & 0\end{pmatrix}, \tag{3.10}$$

which obviously commute with each other, we have

$$\sigma_\pm = \exp(\log\sigma)_\pm \tag{3.11}$$

in coincidence with (3.3). It is also possible to check (3.3) directly.

The behavior at infinity is determined by

$$\begin{aligned}\gamma_+(\xi) &\sim \begin{cases}\sqrt{k}\,\xi^{-1/2}, & \xi\to+\infty,\\ \sqrt{2/k}\,i|\xi|^{1/2}, & \xi\to-\infty,\end{cases}\\ \gamma_-(\xi) &\sim \begin{cases}(i/\sqrt{k})\xi^{1/2}, & \xi\to+\infty,\\ \sqrt{k/2}\,|\xi|^{-1/2}, & \xi\to-\infty.\end{cases}\end{aligned} \tag{3.12}$$

Thus the arguments of $\gamma_\pm$ are bounded and the entries of exponential type in $\sigma_\pm$ behave as

$$\begin{aligned}|\exp[C\cdot\log\gamma_\pm(\xi)]| &= O(\exp\{\operatorname{Re} C\cdot\log|\gamma_\pm(\xi)|\})\\ &= \begin{cases}O(|\xi|^{\mp\operatorname{Re}C/2}), & \xi\to+\infty\\ O(|\xi|^{\pm\operatorname{Re}C/2}), & \xi\to-\infty.\end{cases}\end{aligned} \tag{3.13}$$

Note that this is bounded (the hyperbolic terms oscillate), if and only if $C\in i\mathbb{R}$ holds or equivalently $\lambda > 1$ corresponding to the case that was excluded in (2.3).

4. How to find a generalized factorization. To end up with the inverse of W we try to consider (3.3) as a factorization in the sense of (2.13)—after the lifting process described in (2.6)–(2.8).

It will be seen that this is not possible without an additional modification: the splitting-off of a polynomial matrix in the middle of (3.3). In our opinion this phenomenon is typical of problems with more general boundary or transmission conditions.

Furthermore, we will see that here usually (in contrast to the scalar Dirichlet or Neumann problems) the plus/minus factors correspond to unbounded operators on $L^2(\mathbb{R})^2$, i.e., $\sigma_{0\pm}\in L^\infty(\mathbb{R})^{2\times 2}$ holds only in "exceptional cases," which makes the interpretation in the spirit of singular equations [16] most important.

For a better understanding of the (linear) operator theoretical structure of the factorization, we consider W_0 in (2.7) rather on the whole scale of L^p spaces $L^p(\mathbb{R})^2$, $1 \leqq p \leqq \infty$ (the elements occurring in (2.8) are L^p multipliers [11]), or at least for $1 < p < \infty$ (because of relations to singular integral operators). This, in fact, means that we seek $u_{|\Omega^\pm} \in W^{s,p}(\Omega^\pm)$ with $s - 1/p = \frac{1}{2}$, i.e., $s = \frac{1}{2} + 1/p$, which is physically less important.

It is known that the Fredholm property of W_0 on $L^p(\mathbb{R}_+)^2$ is equivalent to a factorization (2.13) with upper/lower function matrices

$$\begin{aligned} &\sigma_{0-},\ \sigma_{0+}^{-1} \in L^p(\mathbb{R}, \rho^{2/p})^{2\times 2}, \\ &\sigma_{0+},\ \sigma_{0-}^{-1} \in L^q(\mathbb{R}, \rho^{2/q})^{2\times 2}, \qquad 1/p + 1/q = 1 \end{aligned} \tag{4.1}$$

(see [16, V, Thm. 5.2]). Further note that $\varphi(\xi) = |\xi - k|^\mu$, $\xi \in \mathbb{R}$, implies

$$\varphi \in L^p(\mathbb{R}, \rho^{2/p}) \Leftrightarrow \mu p < 1. \tag{4.2}$$

We call $\operatorname{ord} \varphi = \mu$ the *exact order of* φ (at $\pm\infty$), and the order of a matrix function is defined elementwise.

Now we are able to test whether the lifted function theoretic factorization defined by

$$\tilde{\sigma}_{0-}\tilde{\sigma}_{0+} = \begin{pmatrix} t_- & 0 \\ 0 & t_-^{-1} \end{pmatrix} \sigma_- \cdot \sigma_+ \begin{pmatrix} t_+^{-1} & 0 \\ 0 & t_+ \end{pmatrix} \tag{4.3}$$

is a generalized factorization with trivial middle term ($\kappa_1 = \kappa_2 = 0$).

LEMMA 4.1. *Formula* (4.3) *is not a generalized factorization for any* $p \in (1, \infty)$ *and any parameter* $\lambda \in \mathbb{C}$.

Proof. From (3.3), (3.13), and (4.3) we obtain for $0 \neq \lambda \neq 1$ and $\delta = \operatorname{Re} C \in [0, 1]$ (see the Appendix)

$$\begin{aligned} \operatorname{ord} \tilde{\sigma}_{0-} = \operatorname{ord} \tilde{\sigma}_{0+}^{-1} &= \begin{pmatrix} \frac{1}{2}(\delta+1) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(\delta+1) & \frac{1}{2}(\delta-1) \end{pmatrix}, \\ \operatorname{ord} \tilde{\sigma}_{0+} = \operatorname{ord} \tilde{\sigma}_{0-}^{-1} &= \begin{pmatrix} \frac{1}{2}(\delta-1) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(\delta+1) & \frac{1}{2}(\delta+1) \end{pmatrix} \end{aligned} \tag{4.4}$$

with regard to $\operatorname{ord} \det \sigma_\pm = 0$.

Thus (4.1)-(4.2) cannot be satisfied for any p, since $\mu = \frac{1}{2}(\delta+1)$, $1/p > \mu \geqq \frac{1}{2}$ and $1/q > \mu \geqq \frac{1}{2}$ contradicts $1/p + 1/q = 1$.

LEMMA 4.2. *For* $\delta = \operatorname{Re} C = \operatorname{Re}(i/\pi) \log(\sqrt{\lambda}+1)/(\sqrt{\lambda}-1) \in [0, 1)$ *and* $|1/p - \frac{1}{2}| > \delta/2$, *a generalized factorization reads*

$$\begin{aligned} \sigma_0 &= \sigma_{0-} \begin{pmatrix} ((\xi-i)/(\xi+i))^{\kappa_1} & 0 \\ 0 & ((\xi-i)/(\xi+i))^{\kappa_2} \end{pmatrix} \sigma_{0+} \\ &= \begin{cases} \tilde{\sigma}_{0-} \begin{pmatrix} (\xi-i)^{-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} (\xi-i)/(\xi+i) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi+i & 0 \\ 0 & 1 \end{pmatrix} \tilde{\sigma}_{0+}, & p < \dfrac{2}{1+\delta} \\ \tilde{\sigma}_{0-} \begin{pmatrix} 1 & 0 \\ 0 & \xi-i \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (\xi+i)/(\xi-i) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (\xi+i)^{-1} \end{pmatrix} \tilde{\sigma}_{0+}, & p > \dfrac{2}{1-\delta}. \end{cases} \end{aligned} \tag{4.5}$$

Thus the index of W_0 *acting on* $L^p(\mathbb{R}_+)^2$ *is given by*

$$\operatorname{Ind} W_0 = -(\kappa_1 + \kappa_2) = \mp 1 \tag{4.6}$$

for $p<2/(1+\delta)$ and $p>2/(1-\delta)$, respectively, and a left/right inverse is given by

$$W_0^- = A_{0+}^{-1}1_+ \cdot U^{-1}1_+ \cdot A_{0-}^{-1}|_{L^p(\mathbb{R}_+)^2} \tag{4.7}$$

with $A_{0\pm} = F^{-1}\sigma_{0\pm} \cdot F$ and $U = F^{-1}(((\xi-i)/(\xi+i))^{\kappa_j}\delta_{jk}) \cdot F$.

Proof. It is easy to verify the properties (4.1), since all elements of σ_{0-} and σ_{0+} are of order $\frac{1}{2}(\delta-1)$ and $\frac{1}{2}(\delta+1)$. Since ord det $\sigma_{0-} = -1$ in the first case (see (2.8), (3.3), (4.5) (det $\sigma_\pm = $ const!)), we have ord $\sigma_{0-}^{-1} = +1+\frac{1}{2}(\delta-1) = \frac{1}{2}(\delta+1)$ in any place etc., which yields the corresponding properties (4.1) of the inverse matrices. Thus (4.5) represents a generalized factorization and the following conclusions are standard [16], [21].

Remark 4.3. This trick obviously does not work for $1/p \in [\frac{1}{2}(1-\delta), \frac{1}{2}(1+\delta)]$, and in particular not for $p=2$ where Ind $W_0 = 0$ holds.

For $\delta = \operatorname{Re} C = 0$, i.e., $\lambda \in (1, \infty)$, see the Appendix, where only the case $p=2$ is excluded (cf. the remark after (3.13)).

For $\delta = 1$, i.e., $\lambda \in (0, 1)$, we do not have any result so far.

THEOREM 4.4. *For $\delta \in (0, 1]$, and $|1/p - \frac{1}{2}| < \delta/2$, a generalized factorization reads*

$$\sigma_0 = \sigma_{0-} \cdot \sigma_{0+} = \tilde{\sigma}_{0-}\begin{pmatrix} 1 & 0 \\ \xi_1/\sqrt{\lambda} & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -\xi_1/\sqrt{\lambda} & 1 \end{pmatrix}\tilde{\sigma}_{0+}. \tag{4.8}$$

Thus W_0 and W are invertible and

$$W_0^{-1} = A_{0+}^{-1}1_+ \cdot A_{0-}^{-1}|_{L^p(\mathbb{R}_+)^2} \tag{4.9}$$

holds according to this factorization.

Proof. The following cancellation in the asymptotic behavior of the terms in (3.3) becomes most important. Abbreviating the hyperbolic terms there by $c_\pm$, $s_\pm$, and the exponential by $e_\pm$ we obtain

$$\begin{aligned} c_+\xi + s_+t &= e_+(\xi+t) + e_+^{-1}(\xi-t) \\ &= \begin{cases} O(|\xi|^{\delta/2+1}) + O(|\xi|^{\delta/2-1}), & \xi \to +\infty \\ O(|\xi|^{\delta/2-1}) + O(|\xi|^{-\delta/2+1}), & \xi \to -\infty \end{cases} \\ &= O(|\xi|^{1-\delta/2}), \qquad |\xi| \to \infty \end{aligned} \tag{4.10}$$

according to (3.13), $\delta \in (0, 1]$ and

$$\xi - t = \frac{k^2}{\xi+t} = O(\xi^{-1}), \qquad \xi \to +\infty. \tag{4.11}$$

By analogy, there holds

$$c_-t - s_-t = O(|\xi|^{1-\delta/2}), \qquad |\xi| \to \infty \tag{4.12}$$

in contrast to

$$c_+\xi - s_+t,\ c_-t + s_-\xi = O(|\xi|^{1+\delta/2}). \tag{4.13}$$

Thus we obtain the following lifted modified factors and their orders:

$$\begin{aligned} \sigma_{0+} &= \begin{pmatrix} 1 & 0 \\ -\xi/\sqrt{\lambda} & 1 \end{pmatrix}\sigma_+\begin{pmatrix} t_+^{-1} & 0 \\ 0 & t_+ \end{pmatrix} \\ &= (1-\lambda^{-1})^{1/4}\begin{pmatrix} c_+ & -s_+\sqrt{\lambda}/t \\ -c_+\xi/\sqrt{\lambda} - s_+t/\sqrt{\lambda} & s_+\xi/t + c_+ \end{pmatrix}\begin{pmatrix} t_+^{-1} & 0 \\ 0 & t_+ \end{pmatrix}, \\ \operatorname{ord}\sigma_{0+} &= \begin{pmatrix} \frac{1}{2}(\delta-1) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(1-\delta) & \frac{1}{2}(1-\delta) \end{pmatrix} \end{aligned} \tag{4.14}$$

and

$$
\begin{aligned}
\sigma_{0-} &= \begin{pmatrix} t_- & 0 \\ 0 & t_-^{-1} \end{pmatrix} \sigma_- \begin{pmatrix} 1 & 0 \\ \xi/\sqrt{\lambda} & 1 \end{pmatrix} \\
&= (1-\lambda^{-1})^{1/4} \begin{pmatrix} t_- & 0 \\ 0 & t_-^{-1} \end{pmatrix} \begin{pmatrix} c - s_-\xi/t & -s_-\sqrt{\lambda}/t \\ -s_- t/\sqrt{\lambda} + c_-\xi/\sqrt{\lambda} & c_- \end{pmatrix}, \\
\operatorname{ord} \sigma_{0-} &= \begin{pmatrix} \frac{1}{2}(1-\delta) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(1-\delta) & \frac{1}{2}(\delta-1) \end{pmatrix}.
\end{aligned} \tag{4.15}
$$

Now we ask to which $L^p(\mathbb{R}, \rho^{2/p})$ spaces the elements φ of $\sigma_{0\pm}^{\pm 1}$ belong. Obviously,

$$
\operatorname{ord} \varphi = \mu \leqq \max\{\tfrac{1}{2}(1-\delta), \tfrac{1}{2}(\delta-1)\} = \tfrac{1}{2}(1-\delta) \tag{4.16}
$$

holds for $\delta \in (0, 1]$. So p is determined by, see (4.1)-(4.2)

$$
1/p > \tfrac{1}{2}(1-\delta) \tag{4.17}
$$

and the same condition must be satisfied for $1/q = 1 - 1/p$, i.e., $|1/p - \frac{1}{2}| < \delta/2$.

Remark 4.5. The operators $A_{0\pm}^{\pm 1}$ are not bounded for $\delta \neq 1$ but the composition $A_{0-}^{-1} 1_+ \cdot A_{0-}$ is (see [16, V.5]).

COROLLARY 4.6. *In terms of the characteristic parameter $\lambda = ad/bc$ (see Theorem 2.1) we obtain the following alternative conclusions for the $p = 2$ theory of W_0.*

(1) *$\lambda = 0$: σ_0 is triangular. W_0 decomposes into two single Wiener-Hopf equations that are invertible (see [22]).*

(2) *$\lambda = 1$: σ_0 degenerates, and $R(W_0)$ is not closed (for any p).*

(3) *$\lambda \in (1, \infty)$: σ_0 admits a function theoretic factorization, but not a generalized factorization in $L^2(\mathbb{R}_+)^2$ (but in $L^p(\mathbb{R}_+)^2$, $p \neq 2/(1 \pm \delta)$), and W_0 is not Fredholm for $p = 2$.*

(4) *$\lambda \in (0, 1)$: σ_0 admits a bounded factorization due to $\delta = \operatorname{Re} C = 1$, and W_0 is invertible (for every $p \in (1, \infty)$).*

(5) *$\lambda \notin [0, \infty)$: σ_0 admits an unbounded factorization due to $\delta < 1$ with vanishing partial indices, and W_0 is invertible (also for $p \in (2/(1+\delta), 2/(1-\delta))$ and one-sided invertible for p outside the closed interval).*

5. The explicit solution and its asymptotics. We continue with the one-medium case and assume σ to have standard form (3.1), which leads to the one-parameter family of reference problems $\mathcal{P}_\lambda$:

$$
\begin{aligned}
&(\Delta + k^2) u = 0 && \text{in } \Omega^+ \cup \Omega^-, \\
&-2u_0^- = h_0 && \text{on } \mathbb{R}_+, \\
&(1-\lambda^{-1}) u_1^+ - (1+\lambda^{-1}) u_1^- = h_1 && \text{on } \mathbb{R}_+, \\
&u_0^+ - u_0^- = h_0' && \text{on } \mathbb{R}_-, \\
&u_1^+ - u_1^- = h_1' && \text{on } \mathbb{R}_-
\end{aligned} \tag{5.1}
$$

(see (1.1)-(1.2) and (2.4)). Other realizations for the same parameter are available.

COROLLARY 5.1. *$\mathcal{P}_\lambda$ is well-posed (with respect to the space setting in (1.1)-(1.2)), if and only if $\lambda^{-1} \in \mathbb{C} \setminus [0, 1]$ holds. The solution is then given by (1.3), (1.7) where the unique solution of the Wiener-Hopf system (1.8) must be inserted. This reads*

$$
\begin{aligned}
\begin{pmatrix} v_+ \\ w_+ \end{pmatrix} &= W^{-1} \begin{pmatrix} h_0^* \\ h_1^* \end{pmatrix} \\
&= A_+^{-1} 1_+ \cdot A_-^{-1} \begin{pmatrix} l_e h_0^* \\ l_o h_1^* \end{pmatrix}
\end{aligned} \tag{5.2}
$$

where the Fourier symbols $\Phi_\pm$ *of* $A_\pm$ *are represented by*

$$\Phi_+ = \begin{pmatrix} 1 & 0 \\ -\xi/\sqrt{\lambda} & 1 \end{pmatrix} \sigma_+, \qquad \Phi_- = \sigma_- \begin{pmatrix} 1 & 0 \\ \xi/\sqrt{\lambda} & 1 \end{pmatrix} \tag{5.3}$$

(*see* (4.14)-(4.15)). *Again,* $A_\pm$ *are not bounded operators in* $\mathscr{L}(H^{1/2} \times H^{-1/2}, L^2 \times L^2)$ *and* $\mathscr{L}(L^2 \times L^2, H^{1/2} \times H^{-1/2})$, *respectively, for* $\delta \neq 1$. *But there hold* (*despite the unboundedness*)

$$\begin{gathered} A_-^{-1} 1_+ \cdot A_- \in \mathscr{L}(L^2 \times L^2), \\ W^{-1} \in \mathscr{L}(H^{1/2}(\mathbb{R}_+) \times H^{-1/2}(\mathbb{R}_+), \tilde{H}^{1/2}(\mathbb{R}_+) \times \tilde{H}^{-1/2}(\mathbb{R}_+)). \end{gathered} \tag{5.4}$$

THEOREM 5.2. *If the data* h_0^*, h_1^* *in problem* $\mathscr{P}_\lambda$ *are sufficiently smooth and rapidly decreasing* (*like* $e^{-x_1} 1_+(x_1)$), *the singular behavior of the solution is described by*

$$\nabla u(x) \sim \text{const} \cdot |x|^{\delta/2-1}, \qquad |x| \to 0 \tag{5.5}$$

with $\delta = \operatorname{Re}(i/\pi) \log(\sqrt{\lambda}+1)/(\sqrt{\lambda}-1) \in (0, 1]$ *provided* $\lambda \notin [1, \infty)$.

Proof. By standard arguments [5], [6], [17] the singular behavior depends on the maximal order in the symbol matrix of A_+^{-1};

$$\operatorname{ord} \Phi_+^{-1} = \begin{pmatrix} \delta/2 & \delta/2-1 \\ 1-\delta/2 & -\delta/2 \end{pmatrix}. \tag{5.6}$$

So the gradient of the solution behaves like the inverse Fourier transform of $(\xi^2+1)^{1/2(1-\delta/2)}(\xi+i)^{-1}$.

6. On the two-media case. We finish with some remarks about problems $\mathscr{P}$ where the wave numbers k_1, k_2 differ and continue considering the matrices $\sigma(\xi) = \sigma(\xi; k_1, k_2)$ and $\sigma_0(\xi) = \sigma_0(\xi; k_1, k_2, k)$ in (2.4) and (2.8), respectively.

Remark 6.1. If $\mathscr{P}$ is of normal type, $k_1 \neq k_2$ holds and if σ is of Khrapkov type [10]

$$\sigma = \mu_1 R_1 + \mu_2 R_2 \tag{6.1}$$

with scalar functions μ_j and polynomial matrices R_j, then it must be even of triangular form, i.e., the Wiener-Hopf system decomposes. More precisely there holds that (i) both of the main diagonal elements disappear (two isolated boundary value problems), or (ii) they are linear dependent and at least one of the other two disappears (excluded due to an isolated Wiener-Hopf equation with another suitable space setting; see Remark 2.3.3(i) or (iii) where both of the off-diagonal elements disappear (trivial, since σ is constant)).

So the explicit factorization method of § 3 does not help in any case where $k_1 \neq k_2$ although the existence question is completely answered by Corollary 2.2 (with unknown partial indices).

Remark 6.2. The idea to invert W_0 using the fixed point principle [15], [21] works in the case $\lambda \in (0, 1)$ where we have a bounded factorization. Consider $\sigma_0(\xi; k_1, k_2)$ as a perturbation of $\sigma_0(\xi; k)$ with $k = \frac{1}{2}(k_1+k_2)$, $\operatorname{Im} k_j > 0$, and the same coefficients, i.e., the same characteristic parameter λ. Using the factorization (4.8) $\sigma_0 = \sigma_{0-}\sigma_{0+}$ of $\sigma_0(\xi; k)$ we obtain

$$\sigma_0(\xi; k_1, k_2) = \sigma_0(\xi; k)(I + \sigma_\varepsilon(\xi))$$

with $\|\sigma_\varepsilon\|_{L^\infty(\mathbb{R})^{2\times 2}} < 1$ for $|k_1 - k_2| < \delta(\varepsilon)$ and further

$$= \sigma_{0-}(\xi; k)(I + \tilde{\sigma}_\varepsilon(\xi))\sigma_{0+}(\xi; k)$$

with $\|\tilde{\sigma}_\varepsilon\|_{L^\infty(\mathbb{R})^{2\times 2}} < 1$ for $\|k_1 - k_2\| < \tilde{\delta}(1)$.

This yields the (usual) representation

$$W_0^{-1} = F^{-1}\sigma_{0+}^{-1} \cdot F \sum_{j=0}^{\infty} (1_+ \cdot F^{-1}\tilde{\sigma}_\varepsilon(\xi) \cdot F)^j 1_+ \cdot F^{-1}\sigma_{0-}^{-1} \cdot F|_{L^2(\mathbb{R}_+)^2}.$$

Sufficient estimates for k_j can be obtained easily, in contrast to the case $\lambda \in \mathbb{C}\backslash[0, \infty)$ where the factors are unbounded and we only know about the invertibility in a nonspecified neighborhood of W.

Acknowledgments. The author thanks Robert Allan Hurd who was a guest professor at the Technical University of Darmstadt, funded by the DFG in 1986, for many fruitful discussions that initiated this work and that lead to the explicit factorization of the symbol matrices under consideration.

Appendix. The parameter regions. The admissible values of C in (3.3) are described dependent of $\lambda \in \mathbb{C}$. Another choice of the branches leads to the same factorization. Here we have Re $C \in [0, 1]$, Im $C \in (0, \infty)$ or Re $C \in (0, 1)$, Im $C \in (-\infty, 0]$ (Figs. 1-4).

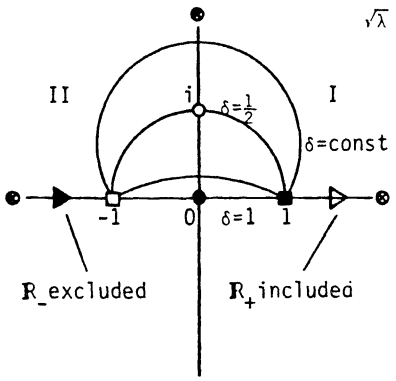


Fig. 1

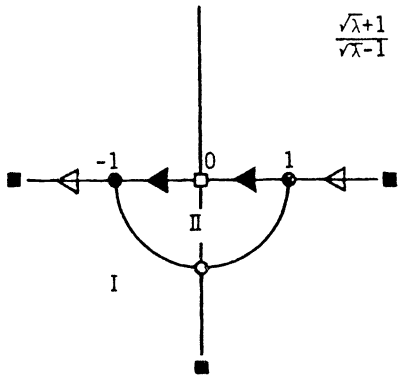


Fig. 2

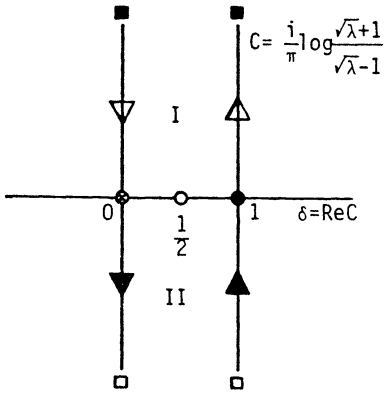


Fig. 3

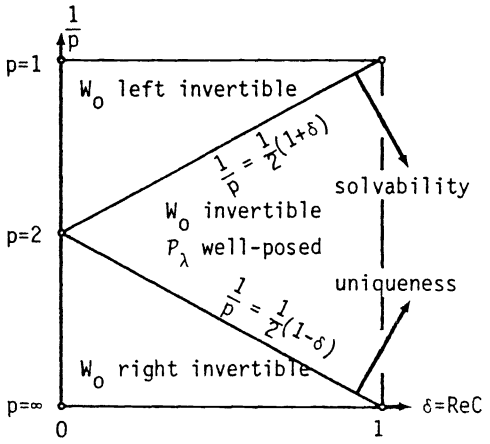


Fig. 4

REFERENCES

[1] K. CLANCEY AND I. GOHBERG, *Factorization of Matrix Functions and Singular Integral Operators*, Birkhäuser, Basel, 1981.

[2] V. G. DANIELE, *On the factorization of Wiener-Hopf matrices in problems solvable with Hurd's method*, IEEE Trans. Antennas and Propagation, 26 (1978), pp. 614-616.

[3] ———, *On the solution of two coupled Wiener-Hopf equations*, SIAM J. Appl. Math., 44 (1984), pp. 667-680.

[4] A. DEVINATZ AND M. SHINBROT, *General Wiener-Hopf operators*, Trans. Amer. Math. Soc., 145 (1969), pp. 467-494.

[5] G. I. ÈSKIN, *Boundary Value Problems for Elliptic Pseudodifferential Equations*, American Mathematical Society, Providence, RI, 1981. (In Russian 1973.)

[6] A. E. HEINS, *Systems of Wiener-Hopf equations and their application to some boundary value problems in electromagnetic theory*, Proc. Sympos. Appl. Math., II (1950), pp. 76-81.

[7] ———, *The Sommerfeld half-plane problem revisited*, II *The factoring of a matrix of analytic functions*, Math. Meth. Appl. Sci., 5 (1983), pp. 14-21.

[8] R. A. HURD, *The Wiener-Hopf-Hilbert method for diffraction problems*, Canad. J. Phys. 54 (1976), pp. 775-780.

[9] ———, *The Explicit Factorization of Wiener-Hopf Matrices*, Preprint 1040, Fachbereich Mathematik, Technische Hochschule Darmstadt, 1987.

[10] A. A. KHRAPKOV, *Certain cases of the elastic equilibrium of an infinite wedge with a non-symmetric notch at the vertex, subjected to concentrated force*, Prikl. Mat. Mekh. 35 (1971), pp. 625-637.

[11] R. LARSEN, *The Multiplier Problem*, Springer-Verlag, Berlin, 1969.

[12] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972. (In French 1968.)

[13] E. LÜNEBURG AND R. A. HURD, *On the diffraction problem of a half-plane with different face impedances*, Canad. J. Phys., 62 (1984), pp. 853-860.

[14] E. MEISTER, *Some multiple-part Wiener-Hopf problems in mathematical physics*, St. Banach Center Publ., 15 (1985), pp. 359-407.

[15] E. MEISTER AND F.-O. SPECK, *Diffraction problems with impedance conditions*, Appl. Anal., 22 (1986), pp. 193-211.

[16] S. G. MIKHLIN AND S. PRÖSSDORF, *Singular Integral Operators*, Springer-Verlag, Berlin, 1986. (In German 1980.)

[17] B. NOBLE, *Methods Based on the Wiener-Hopf Technique for the Solution of Partial Differential Equations*, Pergamon, London, 1958.

[18] A. D. RAWLINS, *The solution of a mixed boundary value problem in the theory of diffraction by a semi-infinite plane*, Proc. Roy. Soc. London Ser. A, 346 (1975), pp. 469-484.

[19] ———, *The explicit Wiener-Hopf factorisation of a special matrix*, Z. Angew. Math. Mech., 61 (1981), pp. 527-528.

[20] A. F. DOS SANTOS AND F. S. TEIXEIRA, *On a class of Wiener-Hopf equations of the first kind in a Sobolev space*, Integral Equations Operator Theory, 10 (1987), pp. 62-81.

[21] F.-O. SPECK, *General Wiener-Hopf Factorization Methods*, Pitman, London, 1985.

[22] ———, *Mixed boundary value problems of the type of Sommerfeld's half-plane problem*, Proc. Roy. Soc. Edinburgh, 104 (1986), pp. 261-277.

[23] W. E. WILLIAMS, *Recognition of some readily "Wiener-Hopf" factorizable matrices*, IMA J. Appl. Math., 32 (1984), pp. 367-378.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 408-414, March 1989

009

GREEN-RIEMANN FUNCTIONS FOR A CLASS OF HYPERBOLIC FOCAL POINT PROBLEMS*

NEZAM IRANIPARAST†

Abstract. Focal point problems such as those in [*Hiroshima Math J.*, 14 (1984), pp. 203-210] in a characteristic triangle have been extended to similar problems for hyperbolic equations of the form $u_{tt}-u_{ss}+q(s,t)u=-\lambda p(s,t)u$. Criteria for the existence of eigenvalues are established by means of Riemann and the generalized Green functions.

Key words. focal point, eigenvalue, Riemann function, Green-Riemann function, Riemann's method, Cauchy problem

AMS(MOS) subject classification. 35L05

1. Introduction. In [5], Kreith has used d'Alembert's method and the theory of positive operators to study certain hyperbolic boundary value problems. In this paper we extend some of the results of [5] to a more general problem using Riemann's method and the theory of positive operators.

Historically, one of the basic assumptions underlying mathematical models of a vibrating string has been the notion of "simultaneous crossing of the axis" [1]. This assumption refers to the phenomenon of the string passing through its equilibrium state, a quite feasible occurrence for free strings. However, for a string under an arbitrary linear restoring force, such a phenomenon is more difficult to establish. Kreith [4] has studied one such formulation, a semidiscrete case of the following problem:

$$u_{tt}-u_{ss}+p(s,t)u=0, \qquad 0<s<L, \quad t>0, \tag{1.1}$$

$$u_s(0,t)=u_s(L,t)=0, \qquad t>0, \tag{1.2}$$

$$u(s,0)=0, \quad u_t(s,0)=g(s), \quad 0\leqq s\leqq L, \tag{1.3}$$

$$u(s,T)=0, \quad T=\text{const.}, \quad 0\leqq s\leqq L. \tag{1.4}$$

In a related paper by Kreith [5], the spatial boundary conditions (1.2) are left out and the differential equation (1.1) is considered over the characteristic triangle

$$R(s,t)=\{(\sigma,\tau): s-(t-\tau)<\sigma<s+(t-\tau), 0<\tau<t\}.$$

Under suitable initial conditions of the type (1.3) the string is then required either to pass through its equilibrium or to be at rest at a single point $s=0$ at time $t=T$, which constitutes a modification of condition (1.4). The "hyperbolic focal point problems" of [5] are then many-degrees-of-freedom analogues of the boundary value problems for $d^2u/dt^2+pu=0$, $(p>0)$.

In this paper we extend the techniques of [5] involving an equation of the type

$$u_{tt}-u_{ss}+\lambda p(s,t)u=0, \quad (s,t) \text{ in } R(0,T),$$

under appropriate initial and time boundary conditions, to the equation

$$u_{tt}-u_{ss}+(q(s,t)+\lambda p(s,t))u=0, \quad (s,t) \text{ in } R(0,T), \tag{1.5}$$

under the same conditions.

* Received by the editors April 20, 1987; accepted for publication (in revised form) June 1, 1988.
† Department of Mathematics, Western Kentucky University, Bowling Green, Kentucky 42101.

In § 2 we use the Green function method of [5] to establish a positive eigenvalue and its corresponding nonnegative eigenfunction for (1.5) when q is a constant. Under the initial and boundary conditions of § 3, the method runs into difficulty (as in [5]) when $q \equiv 0$, but succeeds when $q = c \not\equiv 0$. Under the conditions of § 4, we see that (1.5) can only be transformed into an integral equation. No solution has been established in this case.

2. Right focal points. We apply the technique of [5] to the problem

$$\begin{aligned} &u_{tt} - u_{ss} + q(s,t)u = -\lambda p(s,t)u \quad \text{in } R(0,T), \\ &u_t(s,0) = 0 \qquad\qquad\qquad\qquad\quad \text{in } [-T,T], \\ &u(0,T) = 0. \end{aligned}$$

First the data $u(s,0) = kg(s)$ will be assigned to solve the Cauchy problem

$$\begin{aligned} &u_{tt} - u_{ss} + q(s,t)u = -\lambda p(s,t)u \quad \text{in } R(0,T), \\ &u(s,0) = kg(s), \quad u_t(s,0) = 0 \quad \text{in } [-T,T], \end{aligned} \tag{2.1}$$

and then the condition $u(0,T) = 0$ will be imposed to find k.

By Riemann's method

$$u(s,t) = \frac{1}{2}\left(u(s+t,0) + u(s-t,0) + \int_{s-t}^{s+t} (vu_\tau - uv_\tau)\, d\sigma + \iint_{R(s,t)} -\lambda pvu\, d\sigma\, d\tau \right),$$

therefore (2.1) has the solution

$$u(s,t) = \frac{1}{2}\left(kg(s+t) + kg(s-t) - \int_{s-t}^{s+t} kg(\sigma) v_\tau(\sigma, 0; s, t)\, d\sigma \right.$$
$$\left. - \iint_{R(s,t)} \lambda p(\sigma,\tau) v(\sigma,\tau; s, t) u\, d\sigma\, d\tau \right),$$

where $v(\sigma, \tau; s, t)$ satisfies

$$\begin{aligned} &v_{\tau\tau} - v_{\sigma\sigma} + q(\sigma,\tau)v = 0 \quad \text{in } R(s,t), \\ &v \equiv 1 \quad \text{along } \tau = -\sigma + (s+t), \\ &v \equiv 1 \quad \text{along } \tau = \sigma - (s-t). \end{aligned} \tag{2.2}$$

Hence $u(0,T) = 0$ implies

$$k = \frac{\iint_{R(0,T)} \lambda p(\sigma,\tau) v(\sigma,\tau; 0, T)\, u\, d\sigma\, d\tau}{g(T) + g(-T) - \int_{-T}^{T} g(\sigma) v_\tau(\sigma, 0; 0, T)\, d\sigma},$$

and the solution to

$$\begin{aligned} &u_{tt} - u_{ss} + q(s,t)u = -\lambda p(s,t)u \quad \text{in } R(s,t), \\ &u(s,0) = kg(s), \quad u_t(s,0) = 0 \quad \text{in } [-T,T], \\ &u(0,T) = 0, \end{aligned} \tag{2.3}$$

can be written in the form

$$\frac{1}{\lambda} u(s,t) = \iint_{R(0,T)} G(s,t;\sigma,\tau) p(\sigma,\tau) u\, d\sigma\, d\tau,$$

where

$$G(s,t;\sigma,\tau) = \frac{1}{2} \frac{g(s+t) + g(s-t) - \int_{s-t}^{s+t} g(\sigma) v_\tau(\sigma, 0; s, t)\, d\sigma}{g(T) + g(-T) - \int_{-T}^{T} g(\sigma) v_\tau(\sigma, 0; 0, T)\, d\sigma} v(\sigma,\tau; 0, T) \tag{2.4}$$

for (σ, τ) in $R(0, T)-R(s, t)$,

$$=\frac{1}{2}\frac{g(s+t)+g(s-t)-\int_{s-t}^{s+t} g(\sigma)v_\tau(\sigma, 0; s, t)\,d\sigma}{g(T)+g(-T)-\int_{-T}^{T} g(\sigma)v_\tau(\sigma, 0; 0, T)\,d\sigma}\, v(\sigma, \tau; 0, T)$$

$$-\frac{1}{2}v(\sigma, \tau; s, t)$$

for (σ, τ) in $R(s, t)$.

For $q(s, t)\equiv 0$ the Riemann function $v(\sigma, \tau; s, t)\equiv 1$ and the Green-Riemann function (2.4) reduces to

$$G(s, t; \sigma, \tau)=\frac{1}{2}\frac{g(s+t)+g(s-t)}{g(T)+g(-T)} \quad \text{for } (\sigma, \tau) \text{ in } R(0, T)-R(s, t),$$

$$=\frac{1}{2}\frac{g(s+t)+g(s-t)}{g(T)+g(-T)}-\frac{1}{2} \quad \text{for } (\sigma, \tau) \text{ in } R(s, t).$$

In [5] it has been pointed out that such a Green function is not symmetric but is nonnegative for positive concave g in $[-T, T]$. To show that $G(s, t; \sigma, \tau)$ can be made nonnegative in $R(0, T)$ for $q(s, t)\geqq 0$, we need the following lemmas.

LEMMA 2.1. *Let $q(s, t)\geqq 0$ and $v(\sigma, \tau; s, t)\geqq 0$ in $R(s, t)$. Then $v_\tau(\sigma, \tau; s, t)\geqq 0$ in $R(s, t)$.*

Proof. Choose the change of independent variables $x=\sqrt{2}/2((t-\tau)+(s-\sigma))$ and $y=\sqrt{2}/2((t-\tau)-(s-\sigma))$ to transform (2.2) into

$$v_{xy}+\tfrac{1}{2}q(x, y)v=0, \tag{2.5}$$

$$v(x, 0)=1, \quad x \text{ in } [0, \sqrt{2}T],$$

$$v(0, y)=1, \quad y \text{ in } [0, \sqrt{2}T].$$

Equation (2.5) integrated over $[0, y]$ and $[0, x]$ yields

$$v_x(x, y)=-\frac{1}{2}\int_0^y q(x, \eta)v(x, \eta)\,d\eta\leqq 0,$$

$$v_y(x, y)=-\frac{1}{2}\int_0^x q(\xi, y)v(\xi, y)\,d\xi\leqq 0;$$

therefore

$$v_\tau=-\sqrt{2}/2(v_x+v_y)\geqq 0.$$

LEMMA 2.2. *Let $q(s, t)\equiv C>0$, C be a constant, and ω be the first zero of J_0 the Bessel function of order zero. Then for $T\leqq \omega/\sqrt{C}$, $v(\sigma, \tau; s, t)\geqq 0$ for all (s, t) in $R(0, T)$ and (σ, τ) in $R(s, t)$.*

Proof. With $q\equiv C$, the Riemann function $v(\sigma, \tau; s, t)=J_0(\sqrt{C}\Gamma)$, where $\Gamma=\sqrt{(t-\tau)^2-(s-\sigma)^2}$. The function J_0 is nonnegative for $\sqrt{C}\Gamma\leqq\omega$. This implies that J_0 is nonnegative for all (σ, τ) lying in the region between the two branches of the hyperbola $(t-\tau)^2-(s-\sigma)^2=\omega^2/C$. Therefore $R(s, t)$ will lie in this region, for all (s, t) in $R(0, T)$, if $T\leqq\omega/\sqrt{C}$.

It follows from Lemmas 2.1 and 2.2 that for $q\equiv C>0$ and $T\leqq\omega/\sqrt{C}$ the inequalities $v(\sigma, \tau; s, t)\geqq 0$ and $v_\tau(\sigma, \tau; s, t)\geqq 0$ are satisfied for all (s, t) in $R(0, T)$ and (σ, τ) in $R(s, t)$. Let $T_0<\omega/\sqrt{C}$ be chosen such that $v(\sigma, 0; 0, T_0)\geqq m$ for $-T_0\leqq\sigma\leqq T_0$. Let $v_\tau(\sigma, 0; s, t)\leqq M$ for (s, t) in $R(0, T)$, $s-t\leqq\sigma\leqq s+t$. Assume that g is

positive and concave with a maximum g_0 and minimum $g(T)=g(-T)$ over the interval $[-T, T]$, where T satisfies

$$T \leqq \min\,(g(T)/Mg_0,\ T_0). \tag{2.6}$$

Then with T_0, m, M, g_0, $g(T)$, and T so defined we have the following lemma.

LEMMA 2.3. *Let T be defined by* (2.6) *and let g be a positive continuous function with $g'' \leqq 0$ in $[T, -T]$. Then the values of g in $(-T, T)$ can be chosen so that $G(s, t; \sigma, \tau) \geqq 0$ in $R(0, T)$.*

Proof. For the integrals in the numerator and denominator of the expression for G we have

$$\int_{s-t}^{s+t} g(\sigma)v_\tau(\sigma, 0; s, t)d\sigma < 2Tg_0M \leqq (2g(T)/Mg_0)g_0M = 2g(T),$$

$$\int_{-T}^{T} g(\sigma)v_\tau(\sigma, 0; 0, T)d\sigma < 2Tg_0M \leqq (2g(T)/Mg_0)g_0M = 2g(T).$$

Therefore $g(s+t)+g(s-t)$ can be chosen large enough in $(-T, T)$ so that

$$\frac{g(s+t)+g(s-t)-\int_{s-t}^{s+t} g(\sigma)v_\tau(\sigma, 0; s, t)\,d\sigma}{g(T)+g(-T)-\int_{-T}^{T} g(\sigma)v_\tau(\sigma, 0; 0, T)\,d\sigma} \geqq \frac{1}{m} \geqq \frac{v(\sigma, \tau; s, t)}{v(\sigma, \tau; 0, T)}.$$

Hence $G(s, t; \sigma, \tau) \geqq 0$ for all (s, t) and (σ, τ) in $R(0, T)$.

For $p \geqq 0$, define A to be the operator

$$A[u] = \iint_{R(0,T)} G(s, t; \sigma, \tau)p(\sigma, \tau)u\,d\sigma\,d\tau.$$

Then A is u_0-positive and completely continuous in $L^q(1 < q < \infty)$ due to the boundedness of G and p [3, Chap. 2, pp. 230–232]. When we choose K to be the cone of functions u nonnegative in $R(0, T)$, the operator A maps K into K. The theory of Krasnoselskii [3, Chap. 2] then yields the following theorem.

THEOREM 2.1. *There exists a T and a function g positive in $[-T, T]$ such that for $p \geqq 0$ in $R(0, T)$ and $q \equiv C > 0$ the eigenvalue problem* (2.3) *has a unique eigenvalue $\lambda_g > 0$ and a corresponding unique nontrivial solution that is nonnegative in $R(0, T)$. All other eigenvalues of* (2.3) *satisfy $|\lambda| > \lambda_g$.*

3. Left focal points. An argument similar to the one in § 2 can be applied to the following eigenvalue problem:

$$\begin{aligned} &u_{tt} - u_{ss} + q(s, t)u = -\lambda p(s, t)u \quad \text{in } R(0, T), \\ &u(s, 0) = 0, \quad u_t(s, 0) = kg(s) \quad \text{in } [-T, T], \end{aligned} \tag{3.1}$$

$$u_t(0, T) = 0. \tag{3.2}$$

However, for $q(s, t) \equiv 0$ the technique of § 2 fails. In this case a result has been established in [5] for the existence of a focal point for

$$\begin{aligned} &u_{tt} - u_{ss} + p(s, t)u = 0 \quad \text{in } R(0, T), \\ &u(s, 0) = 0, \quad u_t(s, 0) = kg(s) \quad \text{in } [-T, T], \\ &u_t(0, T) = 0. \end{aligned}$$

Again solve the Cauchy problem (3.1) by the Riemann method to obtain

$$u(s, t) = \frac{1}{2}\left(\int_{s-t}^{s+t} kg(\sigma)v(\sigma, 0; s, t)\,d\sigma - \iint_{R(s,t)} \lambda p(\sigma, \tau)v(\sigma, \tau; s, t)u\,d\sigma\,d\tau\right),$$

and so

$$u_t(s,t)=\frac{1}{2}\Big(k(g(s+t)v(s+t,0;s,t)+g(s-t)v(s-t,0;s,t)$$
$$+\int_{s-t}^{s+t} g(\sigma)v_t(\sigma,0;s,t)\,d\sigma)$$
$$-\lambda\int_0^t (p(-\tau+(s+t),\tau)v(-\tau+(s+t),\tau;s,t)u(-\tau+(s+t),\tau)$$
$$+p(\tau+(s-t),\tau)v(\tau+(s-t),\tau;s,t)u(\tau+(s-t),\tau))\,d\tau$$
$$-\lambda\iint_{R(s,t)} p(\sigma,\tau)v_t(\sigma,\tau;s,t)u(\sigma,\tau)\,d\sigma\,d\tau\Big).$$

Impose condition (3.2) and assume that $p(-\tau+T,\tau)=p(\tau-T,\tau)\equiv 0$, $0\leqq\tau\leqq T$ to obtain

$$k=\frac{\lambda\iint_{R(0,T)} p(\sigma,\tau)v_t(\sigma,\tau;0,T)u\,d\sigma\,d\tau}{g(T)+g(-T)+\int_{-T}^{T} g(\sigma)v_t(\sigma,0;0,T)\,d\sigma}.$$

The solution of (3.1), (3.2) can now be written in the form

$$\frac{1}{\lambda}u(s,t)=\iint_{R(0,T)} G(s,t;\sigma,\tau)p(\sigma,\tau)u\,d\sigma\,d\tau,$$

where

$$G(s,t;\sigma,\tau)$$
$$=\frac{1}{2}\frac{\int_{s-t}^{s+t} g(\sigma)v(\sigma,0;s,t)\,d\sigma}{g(T)+g(-T)+\int_{-T}^{T} g(\sigma)v_t(\sigma,0;0,T)\,d\sigma}\,v_t(\sigma,\tau;0,T)$$
$$\text{for } (\sigma,\tau) \text{ in } R(0,T)-R(s,t),$$
$$=\frac{1}{2}\frac{\int_{s-t}^{s+t} g(\sigma)v(\sigma,0;s,t)\,d\sigma}{g(T)+g(-T)+\int_{-T}^{T} g(\sigma)v_t(\sigma,0;0,T)\,d\sigma}$$
$$\cdot v_t(\sigma,\tau;0,T)-\frac{1}{2}v(\sigma,\tau;s,t)\quad \text{for } (\sigma,\tau) \text{ in } R(s,t).$$

As in § 2, for $q\equiv C>0$ the function $v(\sigma,\tau;s,t)=J_0(\sqrt{C}\Gamma)$ and we have the following lemma.

LEMMA 3.1. *Let ω be the first zero of J_0. Then for $T\leqq\omega/\sqrt{C}$ the Riemann function $v(\sigma,\tau;s,t)=J_0(\sqrt{C}\Gamma)$ satisfies $v_t(\sigma,\tau;0,T)\leqq 0$ in $R(0,T)$.*

Proof. Note that for $\Gamma=\sqrt{(T-\tau)^2-(\sigma)^2}$

$$v_t(\sigma,\tau;0,T)=(\sqrt{C}(T-\tau)/\Gamma)\sum_{k=1}^{k=\infty}((-1)^k k/(k!)^2)(\sqrt{C}\Gamma/2)^{2k-1}.$$

To have $v_t(\sigma,\tau;0,T)\leqq 0$ in $R(0,T)$ we let

$$(k/(k!)^2)(\sqrt{C}\Gamma/2)^{2k-1}\geqq((k+1)/((k+1)!)^2)(\sqrt{C}\Gamma/2)^{2k+1}\quad \text{for } k=1,2,3,\cdots,$$

which implies that Γ must satisfy $\Gamma\leqq\sqrt{8}/\sqrt{C}$, i.e., $(T-\tau)^2-(\sigma)^2\leqq 8/C$. The characteristic triangle $R(0,T)$ now lies completely between the two branches of the hyperbola $(T-\tau)^2-(\sigma)^2=8/C$ if $T\leqq\sqrt{8}/\sqrt{C}$. But T already satisfies $T\leqq\omega/\sqrt{C}<\sqrt{8}/\sqrt{C}$, so we are done.

Let g be positive in $[-T, T]$ and $C > 4\omega^2$. Then $T < \frac{1}{2}$ and the values of g at the endpoints of $[-T, T]$ can be chosen large enough so that

$$g(T) + g(-T) + \int_{-T}^{T} g(\sigma) v_t(\sigma, 0; 0, T)\, d\sigma > 0. \tag{3.3}$$

Now by Lemmas 2.2 and 3.1, $G(s, t; \sigma, \tau) \leqq 0$ in $R(0, T)$. If p is chosen nonpositive in $R(0, T)$, then the operator A defined by

$$A[u] = \iint_{R(0,T)} G(s, t; \sigma, \tau) p(\sigma, \tau) u \, d\sigma \, d\tau,$$

is a positive operator on the cone K of functions u nonnegative in $R(0, T)$. Once more Krasnoselskii's theory [3] implies the following theorem.

THEOREM 3.1. *Let $q(s, t) \equiv C > 4\omega^2$ and $T \leqq \omega/\sqrt{C}$ where ω is the first zero of the Bessel function J_0 of order zero. Let $p(s, t) \leqq 0$ in $R(0, T)$ with $p \not\equiv 0$ along the characteristic boundaries of $R(0, T)$. Finally let $g(s) > 0$ in $[-T, T]$ satisfy* (3.3). *Then there is a unique eigenvalue $\lambda_g > 0$ to which there corresponds a unique nontrivial solution of* (3.1), (3.2) *which is nonnegative in $R(0, T)$. All other eigenvalues of* (3.1), (3.2) *satisfy $|\lambda| > \lambda_g$.*

4. Conjugate points. It appears that even the introduction of an extra function q to the conjugate point problem of [5] does not help resolve the establishment of a conjugate point. By the method of the last two sections the problem

$$\begin{aligned} &u_{tt} - u_{ss} + q(s, t)u = -\lambda p(s, t)u \quad \text{in } R(0, T), \\ &u(s, 0) = 0, \quad u_t(s, 0) = kg(s) \quad \text{in } [-T, T], \\ &u(0, T) = 0, \end{aligned} \tag{4.1}$$

can be written in the equivalent form

$$\frac{1}{\lambda} u(s, t) = \iint_{R(0,T)} G(s, t; \sigma, \tau) p(\sigma, \tau) u \, d\sigma \, d\tau,$$

where

$$G(s, t; \sigma, \tau) = \frac{1}{2} \frac{\int_{s-t}^{s+t} g(\sigma) v(\sigma, 0; s, t)\, d\sigma}{\int_{-T}^{T} g(\sigma) v(\sigma, 0; 0, T)\, d\sigma} v(\sigma, \tau; 0, T)$$

for (σ, τ) in $R(0, T) - R(s, t)$,

$$= \frac{1}{2} \frac{\int_{s+t}^{s+t} g(\sigma) v(\sigma, 0; s, t)\, d\sigma}{\int_{-T}^{T} g(\sigma) v(\sigma, 0; 0, T)\, d\sigma} v(\sigma, \tau; 0, T) - \frac{1}{2} v(\sigma, \tau; s, t)$$

for (σ, τ) in $R(s, t)$.

Nevertheless, the Green-Riemann function $G(s, t; \sigma, \tau)$ cannot be made nonnegative or nonpositive in $R(0, T)$. However, the conjugate point problem (4.1), in a slightly different form, has recently been studied by Kreith [6] and Haws [2]. Their method involves construction of a symmetric Green function. The results of [6] and [2] show the existence of such conjugate points.

REFERENCES

[1] J. CANNON AND S. OSTROVSKY, *The Evolution of Dynamics, Vibration Theory from* 1687 *to* 1742, Springer-Verlag, New York, 1981.

[2] L. D. HAWS, *The construction of symmetric Green's functions for a class of hyperbolic boundary value problems*, Ph.D. thesis, 1987.
[3] M. A. KRASNOSELSKII, *Positive Solutions of Operator Equations*, P. Noordhoff, Groningen, 1964.
[4] K. KREITH, *Uniform zeros for beaded strings*, in Proc. Equadiff, 6, Brno, Czechoslovakia, 1985, pp. 141-148.
[5] ———, *A class of hyperbolic focal point problems*, Hiroshima Math. J., 14 (1984), pp. 203-210.
[6] ———, *Symmetric Green's functions for a class of* CIV *boundary value problems*, Canad. Math. Bull., to appear.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 415-429, March 1989

010

THE REGULARITY OF SOLUTIONS OF NONLINEAR WAVE EQUATIONS*

BERNARD MARSHALL†

Abstract. For the wave equation with lower-order nonlinearities a regularity theorem is given which ensures that solutions in certain L^q are in fact smooth solutions to the equation. For equations with energy estimates the existence of global classical solutions in lower dimensions is obtained.

Key words. nonlinear, wave equation, regularity

AMS(MOS) subject classifications. 35L70, 35L15

1. Introduction. The purpose of this paper is to study the regularity of solutions to the Cauchy problem for equations of the form

$$u_{tt}+(-\Delta)^m u+au+f(u)=g \tag{1}$$

in $R_+\times R^n$ with initial data

$$u(0,x)=u_0(x), \qquad u_t(0,x)=u_1(x) \tag{2}$$

where $m\geqq 1$, $\partial^\alpha=\prod(\partial/\partial x_i)^{\alpha_i}$, $|\alpha|=\sum|\alpha_i|$, and $a\geqq 0$. The nonlinear term $f(u)$ is described in § 3. The results we prove apply also to the analogous class associated to the Schrödinger equation:

$$iu_t+(-\Delta)^m u+f(u)=g \tag{3}$$

with initial data $u(0,x)=u_0(x)$.

The existence of classical (C^2) solutions, particularly of (1), has been studied by many authors. The results, however, all have some restriction on the dimension. In the case $m=1$, Jörgens [4] proved the existence of classical solutions if $n=3$. This was extended to $n<10$ by Brenner and von Wahl in [2]. In the case $m\geqq 2$ these solutions have been proven to exist by Pecher [7] and Narasaki [6], who extended this to $n\leqq 6(m-k)+4$. Similar results have been obtained for the Schrödinger equation (3). For example, Tsutsumi and Hayashi [10] have shown the existence of classical solutions when $m=1$ and $n\leqq 9$.

In this paper we present a regularity theorem, valid in all dimensions, which states roughly that a solution is smooth if it contains a critical amount k_0 of differentiability. Since there is no positivity condition imposed on $f(u)$ this theorem applies also to solutions that eventually blow-up. If a positivity condition is added then the existence of global classical solutions in lower dimensions is obtained.

2. The linear estimates. Let $a\geqq 0$. The solution of the Cauchy problem for

$$u_{tt}+(-\Delta)^m u+au=0, \qquad (t,x)\in R_+\times R^n$$

can be written in terms of Fourier multipler transformations with multipliers

$$\cos(t\sqrt{a+|\xi|^{2m}}) \quad \text{and} \quad (a+|\xi|^{2m})^{-1/2}\sin(t\sqrt{a+|\xi|^{2m}}).$$

Denote the corresponding transformations by K_t^C and K_t, respectively. Let $L_k^p(R^n)=(I-\Delta)^{-k/2}(L^p(R^n))$ with norm $|f|_{k,p}\equiv|(I-\Delta)^{k/2}f|_{L^p}$. If k is a nonnegative integer, then an equivalent norm is $|f|_p+\sum_{|\alpha|=k}|\partial^\alpha f|_p$. See [9].

* Received by the editors January 19, 1988, and accepted for publication (in revised form) July 5, 1988.
† Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 2K6, Canada.
This research was supported by National Science and Engineering Research Council of Canada grant A8917.

The following lemma summarizes work by previous authors on these transformations.

THEOREM 1. *Suppose* $p^{-1}+q^{-1}=1$ *and* $0 \leqq A \leqq m$.

(i) *If* $A+mn|1/2-1/p|<m$ *then* K_t^C *is a bounded linear transformation from* $L^p(R^n)$ *to* $L^p_{A-m}(R^n)$ *with norm* $\tilde{\sigma}(t) \leqq C(1+t)^M$, *where* $M=M(n)$. *Similarly* K_t *is a bounded linear transformation from* $L^p(R^n)$ *to* $L^p_A(R^n)$ *with norm* $\sigma(t) \leqq C(1+t)^M$.

(ii) *Assume* $1<p \leqq 2$. *If* $m=1$, *suppose that* $A+(n+1)|1/p-1/2| \leqq 1$. *If* $m>1$, *suppose that* $A+2n|1/p-1/2|<2m$. *Then* K_t *is a bounded linear operator from* $L^p(R^n)$ *to* $L^q_A(R^n)$ *with norm* $\kappa(t)$ *such that* $\int_0^t \kappa(s)\,ds<\infty$ *for all* $t<\infty$.

Proof. (i) Let $\eta(x)$ be a C^∞ function of R^n such that $\eta(x)=0$ if $|x| \leqq 1$ and $\eta(x)=1$ if $|x| \geqq 2$. The multiplier

$$(1-\eta(t^{1/m}\xi))\cos(t\sqrt{a+|\xi|^{2m}})(1+|\xi|^{2m})^{(A-m)/2m}$$

determines a transformation on L^p with norm $\leqq C(1+t)^M$ by the Mikhlin-Hörmander multiplier theorem [3 Thm. 2.5]). The remaining part of the multiplier for K_t^C consists of two terms of the form

$$\frac{\eta(t^{1/m}\xi)\,e^{it\xi_a}}{\xi_*^k}=\frac{\eta(t^{1/m}\xi)\,e^{it|\xi|^m}}{\xi_t^k}\left\{\left(\frac{\xi_t}{\xi_*}\right)^k\right\}\{e^{it(\xi_a-|\xi|^m)}\},$$

where $\xi_t=\sqrt{1+t^2|\xi|^{2m}}$, $\xi_a=\sqrt{a+|\xi|^{2m}}$, $\xi_*=\sqrt{1+|\xi|^{2m}}$, and $k=(m-A)/m$. Each of the multipliers in $\{\ \}$ brackets is bounded on all L^p, $1<p<\infty$, with norms $\leqq C(1+t)^k \leqq C(1+t)$ and $\leqq C(1+t)^M$ by the Mikhlin-Hörmander theorem. A simple change of variables shows that the first multiplier has norm independent of t. But $\eta(\xi)\exp(i|\xi|^m)(1+|\xi|^{2m})^{-k/2}$ is a multiplier on $L^p(R^n)$ if $|1/2-1/p|<k/n=(m-A)/2mn$. If $n=1$, $m=1$, $A=0$ then it is a multiplier for $1<p<\infty$. See [9], p. 113. The proof for K_t is the same. In fact in this case the norm goes to zero as $t \to 0$.

(ii) In this case Pecher [9] has shown that

$$\exp(it\sqrt{a+|\xi|^{2m}})(a+|\xi|^{2m})^{-1/2}(1+|\xi|^2)^{A/2}$$

defines a bounded transformation from $L^p(R^n)$ to $L^q(R^n)$ with norm $\leqq Ct^\rho(1+t)^M$, where $\rho=(n(1-2/p)+m-A)/m$ and $\rho \leqq 0$. If $m=1$, then $1/p \leqq 1/2+(m-A)/(mn+m)$. This shows that $(I-\Delta)^{A/2}K_t$ is bounded with norm $\leqq \kappa(t) \leqq Ct^\rho(1+t)^M$. From the Plancherel theorem this operator is also bounded for L^2 to L^2 with norm $\leqq Ct^{(m-A)/m}$. If $m-A<n$ then $\rho<0$ when $p=1$ and so $(I-\Delta)^{A/2}K_t$ is also bounded from L^p to L^q for p arbitrarily close to one. This shows that the operator is bounded for all $1<p \leqq 2$ and removes the restriction $\rho \leqq 0$ when $m>1$. On the other hand, if $m-A \geqq n$, then the multiplier can be written as

$$(1+|\xi|^2)^{-(m-A)/2}\left\{\frac{\sin(t\sqrt{a+|\xi|^{2m}})}{\sqrt{a+|\xi|^{2m}}}(1+|\xi|^2)^{m/2}\right\}(1+|\xi|^2)^{-(m-A)/2}.$$

The combination of Sobolev's theorem, the Plancherel theorem, and Sobolev's theorem shows that this is bounded from $L^p \to L^2 \to L^2 \to L^q$ with norm $\leqq C$ for all $1<p \leqq 2$.

The only restriction that now remains on p is when $m=1$. It is, however, necessary that $\kappa(t)$ be locally integrable. For this we need $\rho>-1$. For $m>1$ this is equivalent to

$$A+2n\left(\frac{1}{p}-\frac{1}{2}\right)<2m. \tag{4}$$

When $m=1$ this is weaker than the other condition:

$$A+(n+1)\left(\frac{1}{p}-\frac{1}{2}\right) \leqq 1. \tag{5}$$

For the application of Theorem 1, (4) and (5) can be rewritten as

$$2 \leqq q \leqq 2(n+1)/(n-1+2A) \quad \text{if } m=1 \tag{6}$$

$$2 \leqq q < 2n/(n-2m+A) \quad \text{if } m>1. \tag{7}$$

If the denominator is zero or negative this is interpreted as $2 \leqq q < \infty$.

In Example 1 below, the first part of Theorem 1 will be used. This imposes a stronger condition

$$2 \leqq q < 2mn/(mn-2m+2A) \quad \text{if } m>1. \tag{8}$$

For the Schrödinger equation

$$iu_t + (-\Delta)^m = 0$$

the multiplier is $\exp(it|\xi|^{2m})$. Let K_t^S denote the corresponding transformation.

THEOREM 2. *Assume* $1<p\leqq 2$ *and* $p^{-1}+q^{-1}$. *If* $n(1/p-1/2)<m$ *then* K_t^S *is a bounded linear transformation from* $L^p(R^n)$ *to* $L^q(R^n)$ *with norm* $\kappa_S(t)$ *satisfying* $\int_0^t \kappa_S(s)\,ds<\infty$ *for all* $t<\infty$.

The proof of Theorem 2 is similar to that of Theorem 1. For the Schrödinger equation the condition on q is

$$2 \leqq q < 2n/(n-2m) \tag{9}$$

if $2m<n$ and $2\leqq q<\infty$ if $2m \geqq n$.

3. The nonlinear terms. Several types of nonlinearity will be considered for (1). The important information is summarized in the parameters k_0, γ, q, T and two functions $\tilde{\kappa}$ and ω. The basic property is that for all $k \geqq k_0$ there exist a locally integrable function $\tilde{\kappa}$ and a continuous function ω on $[0, T]$ such that

$$|K_t(f(u)(s))|_{k+1,q} \leqq \tilde{\kappa}(t)\omega(|u(s)|_{k,q})|u(s)|_{k+1,q} \tag{10}$$

for s, $t \in [0, T]$. This inequality together with Gronwall's inequality will allow us to show that a solution u in L_k^q is actually in L_{k+1}^q, and so on. The particular values of q, $\tilde{\kappa}$, and ω will not be important in the final result. We will concentrate therefore more on k_0, γ, and T.

Example 1. Consider (1). Let $k_0 \geqq A$ and

$$\begin{aligned} &1 \leqq \gamma \leqq 1+4n(1-A)/(n(n-1+2A)-2(n+1)(k_0-A)) \quad \text{if } m=1 \\ &1 \leqq \gamma < 1+4(m-A)/(mn-2m+2A-2m(k_0-A)) \quad \text{if } m>1. \end{aligned} \tag{11}$$

Let u be a function on $[0, T]$ with values in $L_A^q(R^n)$, where q is a value satisfying (6), (8) and $q \leqq \gamma+1$. The function $u(t)(x)$ will be written as $u(t, x)$. Let N be the number of multi-indices α such that $|\alpha| \leqq A$: $N = 1+n+\cdots n^A$. The nonlinear terms will be assumed to be of the following form

$$f(u)(t, x) = \psi(t, x, u(t, x), \cdots, \partial^\alpha u(t, x), \cdots),$$

where ψ is a real-valued C^∞ function on $[0, T]\times R^{n+N}$. We place the following growth condition on the function ψ:

$$|\partial_t^b \partial_x^\delta \partial_z^\alpha \omega(t, x, z)| \leqq C(1+|z|)^{\max(\gamma-|\alpha|,0)} \tag{12}$$

for all $(t, x, z) \in [0, T]\times R^{n+N}$ and $(b, \delta, \alpha) \in Z_0 \times Z_0^{n+N}$. Also assume that $\partial_x^\delta \psi(t, x, 0) = 0$ for all $\delta \in Z_0^n$, $(t, x) \in [0, T]\times R^n$. Thus $|\partial_x^\delta \psi(t, x, z)| \leqq C|z|$.

For simplicity ψ is assumed to be a real-valued function on $[0, T]\times R^{n+N}$, but the proofs are the same if ψ is a complex-valued function on $[0, T]\times R^n\times C^N$. In fact, for the Schrödinger equation it is necessary to consider such complex-valued nonlinearities.

Let $\eta\in C^\infty(R^n)$ be such that $\eta(z)=0$ if $|z|\leqq 1$ and $\eta(z)=1$ if $|z|\geqq 2$. Define $\psi_0(t, x, z)=(1-\eta(z))\psi(t, x, z)$, $\psi_1=\eta\psi$, and

$$f_i(u)(t, x)=\psi_i(t, x, \cdots, \partial^\alpha u(t, x), \cdots) \tag{13}$$

for $i=0, 1$. Now ψ_0 and its derivatives are bounded, and ψ_1 satisfies

$$|\partial_t^b\partial_x^\delta\partial_z^\alpha\psi_1(t, x, z)|\leqq C|z|^{\gamma-|\alpha|} \quad \text{if } |\alpha|\leqq\gamma. \tag{14}$$

Let $\beta\in Z^n$. We now describe the decomposition of $\partial^\beta(f(u))$ used in the proof of (9). By differentiating (13) we see that

$$\partial^\beta(f_j(u)(t))=\sum_\delta\sum_\alpha(\partial_x^\delta\partial_z^\alpha\psi_j)(u)(t)\sum_\nu c_\nu\prod_{i=1}^{|\alpha|}\partial^{\nu_i}u(t), \tag{15}$$

where $|\nu_i|\geqq 1$, $|\beta|\leqq|\delta|+\sum|\nu_i|$. We may assume that $|\nu_1|\geqq|\nu_2|\geqq\cdots$. Define

$$\begin{aligned}\phi_1(t)&=\partial^\beta(f_1(u)(t))+\sum_\delta\sum_{|\alpha|>\gamma}(\partial_x^\delta\partial_z^\alpha\psi_0)(u)(t)\sum_\nu c_\nu\prod_{i=1}^{|\alpha|}\partial^{\nu_i}u(t)\\ \phi_{j+2}(t)&=\sum_\delta\sum_{|\alpha|=j}(\partial_x^\delta\partial_z^\alpha\psi_0)(u)(t)\sum_\nu c_\nu\prod_{i=1}^{|\alpha|}\partial^{\nu_i}u(t)\end{aligned} \tag{16}$$

for $0\leqq j<\gamma$. Also $p_1=p$, $p_2=q$, $p_{j+2}=q/j$ for $1\leqq j<\gamma$.

LEMMA 1. *Let k_0, q, γ be as described above. If $(k-A)^2+n(1-2/q)\geqq(k-A)n/q$, $k\in Z$ and $k\geqq k_0$ then there exists a continuous function ω_1 on R such that*

$$|f_1(u)(t)|_{k+1-A,p}\leqq\omega_1(|u(t)|_{k,q})|u(t)|_{k+1,q} \tag{17}$$

for all $u(t)\in L^q_{k+1}(R^n)$, $t\in[0, T]$.

Proof. If $|\beta|=k-A+1$ then $\partial^\beta(f_1(u)(t))$ has an expansion as in (15). First consider the terms in (15) where $|\alpha|>\gamma$. In this case $\partial_x^\delta\partial_z^\alpha\psi_1\in L^\infty$, and so this function can be ignored. We wish to apply Hölder's and Sobolev's inequalities to get

$$\left|\prod_{i=1}^{|\alpha|}\partial^{\nu_i}u(t)\right|_p\leqq\prod_{i=1}^{|\alpha|}|\partial^{\nu_i}u(t)|_{p/a_i}\leqq|u(t)|_{k,q}^{|\alpha|-1}|u(t)|_{k+1,q}. \tag{18}$$

To do this we must find $a_i\in[0, 1]$ such that $\sum a_i=1$ and

$$\begin{aligned}\left(\frac{1}{q}-\frac{k+1-|\nu_i|}{n}\right)^+&\leqq\frac{a_i}{p}\leqq\frac{1}{q}, \qquad i=1\\ \left(\frac{1}{q}-\frac{k-|\nu_i|}{n}\right)^+&\leqq\frac{a_i}{p}\leqq\frac{1}{q}, \qquad i\geqq 2,\end{aligned}$$

where $\sum|\nu_i|\leqq(k-A+1)(A+1)$ and $s^+=\max\{s, 0\}$. These inequalities can be written equivalently as

$$\begin{aligned}\frac{p}{n}\left(\frac{n}{q}+|\nu_i|-k-1\right)^+&\leqq a_i\leqq\frac{p}{q}, \qquad i=1\\ \frac{p}{n}\left(\frac{n}{q}+|\nu_i|-k\right)^+&\leqq a_i\leqq\frac{p}{q}, \qquad i\geqq 2.\end{aligned} \tag{19}$$

The sum over $i=1,\cdots,|\alpha|$ of the right-hand terms is $|\alpha|p/q \geqq |\alpha|/\gamma > 1$. If we can show that the sum of the left-hand terms is less than or equal to one then a suitable collection $\{a_i\}$ must exist.

Let

$$S=\left(\frac{n}{q}+|\nu_1|-K-1\right)^+ + \sum_{i=2}^{|\alpha|}\left(\frac{n}{q}+|\nu_i|-k\right)^+.$$

Consider the effect on S of differentiating in (15). Differentiating $\partial^{\nu_i}u$ adds at most one to S while differentiating ψ_1 can add as much as $A+1$ because of the chain rule. If $(n/q)+(A+1)-k<1$ (that is, $k>A+n/q$), then S is largest when $|\alpha|=1$ and $|\nu_1|=k+1$. In this case

$$S\leqq\left(\frac{n}{q}+k+1-k-1\right)=\frac{n}{q}$$

and the sum of the left-hand sides of (19) is bounded by $p/q\leqq 1$. On the other hand, if $k\leqq A+n/q$ when S is largest for terms where $|\alpha|=k-A+1$ and $|\nu_i|=A+1$. In this case

$$\begin{aligned} S &\leqq\left(\frac{n}{q}+A+1-k-1\right)+\sum_{i=2}^{k-A+1}\left(\frac{n}{q}+A+1-k\right) \\ &=(k-A+1)\left(\frac{n}{q}+A+1-k\right)-1=-(k-A)^2+(k-A)\frac{n}{q}+\frac{n}{q}. \end{aligned} \tag{20}$$

Thus the sum of the left-hand sides of (19) is bounded by

$$\frac{p}{n}\left\{-(k-A)^2+(k-A)\frac{n}{q}+\frac{n}{q}\right\}.$$

The condition on k given in the statement of the lemma guarantees that this is less than or equal to one. This completes the justification of (18).

The proof for $|\alpha|\leqq\gamma$ is similar except that in this case the functions $\partial_x^\delta\partial_z^\alpha\psi_1$ are no longer bounded. The terms in (15) are bounded by

$$U(t)=\left(\sum_{|\eta|\leqq A}|\partial^\eta u(t)|\right)^{\gamma-|\alpha|}\prod_{i=1}^{|\alpha|}|\partial^{\nu_i}u(t)| \tag{21}$$

Hölder's and Sobolev's inequalities give

$$\begin{aligned} |U(t)|_p &\leqq\left(\sum_{|\eta|\leqq A}|\partial^\eta u(t)|\right)_{p/a_0}^{\gamma-|\alpha|}\prod_{i=1}^{|\alpha|}|\partial^{\nu_i}u(t)|_{p/a_i} \\ &\leqq|u(t)|_{k,q}^{\gamma-1}|u(t)|_{k+1,q}, \end{aligned}$$

assuming that we can find constants $a_i\in[0,1]$ such that $a_0(\gamma-|\alpha|)+\sum_{i=1}^{|\alpha|}a_i=1$ and

$$\left(\frac{1}{q}-\frac{k-A}{n}\right)^+\leqq\frac{a_0}{p}\leqq\frac{1}{q}, \qquad i=0$$

$$\left(\frac{1}{q}-\frac{k+1-|\nu_i|}{n}\right)^+\leqq\frac{a_1}{p}\leqq\frac{1}{q}, \qquad i=1$$

$$\left(\frac{1}{q}-\frac{k-|\nu_i|}{n}\right)^+\leqq\frac{a_i}{p}\leqq\frac{1}{q}, \qquad i\geqq 2.$$

Equivalently,

$$\frac{p}{n}\left(\frac{n}{q}-k+A\right)^{+} \leqq a_0 \leqq \frac{p}{q}$$

$$\frac{p}{n}\left(\frac{n}{q}+|\nu_1|-k-1\right)^{+} \leqq a_1 \leqq \frac{p}{q}$$

$$\frac{p}{n}\left(\frac{n}{q}+|\nu_i|-k\right)^{+} \leqq a_i \leqq \frac{p}{q}.$$

The weighted sum of the right-hand sides is $\gamma p/q \geqq 1$. For the left-hand sides consider

$$S=(\gamma-|\alpha|)\left(\frac{n}{q}-k+A\right)^{+}+\left(\frac{n}{q}+|\nu_1|-k-1\right)^{+}+\sum_{i=2}^{|\alpha|}\left(\frac{n}{q}+|\nu_i|-k\right)^{+}.$$

If $k-A \geqq n/q$ the maximum of S occurs when $|\alpha|=1$ and $|\nu_1|=A+|\beta|=k+1$. Then

$$S \leqq 0+\left(\frac{n}{q}+|\nu_1|-k-1\right)^{+}=\frac{n}{q} \leqq \frac{n}{p}$$

If $k-A<n/q$ then $|\nu_i| \leqq A+v_i$ where $1 \leqq v_i$ and $\sum v_i \leqq |\beta|=k+1-A$. Then

$$S \leqq (\gamma-|\alpha|)\left(\frac{n}{q}-k+A\right)+\left(\frac{n}{q}+A+v_1-k-1\right)+\sum_{i=2}^{|\alpha|}\left(\frac{n}{q}+v_i-k+A\right)$$

$$=\gamma\frac{n}{q}-(k-A)(\gamma-1)+\sum_{i=1}^{|\alpha|} v_i+A-k-1 \leqq \gamma\frac{n}{q}-(k-A)(\gamma-1).$$

Therefore to ensure that $S \leqq n/p=n-n/q$ we require

$$\gamma n/q \leqq n-n/q+(k-A)(\gamma-1).$$

In terms of γ this is equivalent to

$$\gamma \leqq 1+n(q-2)/(n-(k-A)q). \tag{22}$$

Since the right-hand side is an increasing function of q for $(k-A) \leqq n/2$, we substitute the upper bounds for q given in (6) and (8). Since $k \geqq k_0$, this gives (11).

The proof of the lemma is concluded by summing over $|\beta|=k-A+1$. □

LEMMA 2. *If k_0 is a constant such that $k_0 \geqq A$ and $(k-A)^2+n(1-2/q) \geqq (k-A)n/q$ for all $k \geqq k_0$ then estimate* (10) *holds for all $k \geqq k_0$.*

Proof. Let $|\beta|=k+1-A$. Lemma 1 gives the appropriate estimate for $f_1(u)$. Next we prove the corresponding estimate for $\phi_j(t)$, as defined in (16). The proof of Lemma 1 shows that $|\phi_1(t)|_p$ is bounded by the right-hand side of (17).

For $\phi_2(t)$ we use the fact that

$$|(\partial_x^{\delta}\psi_0)(u)(t)| \leqq C \sum_{|\eta| \leqq A} |\partial_x^{\eta} u(t)| \quad \text{pointwise in } x,$$

because $|(\partial_x^{\delta}\psi_0)(t, x, z)| \leqq C|z|$. Hence

$$|\phi_2(t)|_q \leqq |u(t)|_{A,q} \leqq |u(t)|_{k,q}.$$

For each $1 \leqq j<\gamma$, $p_{j+2}=q/j$ and the derivatives of ψ_0 are bounded. Therefore by Hölder's inequality

$$|\phi_{j+2}(t)|_{q/j} \leqq C \prod_{i=1}^{j} |\partial^{\nu_i} u(t)|_q \leqq C|u(t)|_{k,q}^{j-1}|u(t)|_{k+1,q}.$$

By Theorem 1, K_t is bounded from L^p to L^q_A and from L^q to L^q_A with norms $\kappa(t)$ and $\sigma(t)$. Therefore by interpolation K_t is bounded from L^{p_j} to L^q with norm $\leqq \kappa(t)+\sigma(t)=\tilde{\kappa}(t)$. Using Lemma 1 and the estimates for the ϕ_j gives the inequality (10). As explained in (8) the condition on γ is given in (11).

The other property of $f(u)$ that is used is the observation from (14) that differentiation with respect to t does not change any of the estimates. This is used later in Lemma 9.

Example 2. The condition on k_0 can be relaxed if we consider fractional orders of integration. The use of Besov spaces to accomplish this appears in Brenner [1] and Brenner and von Wahl [2]. The cost of this extra freedom in k, however, is that in (12) we must add the condition

$$|\partial_t^b \partial_x^\sigma \partial_z^\alpha \psi(t, x, z)| \leqq C(1+|z|)^{\gamma-|\alpha|} \quad \text{if } |\alpha| < \gamma+1. \tag{23}$$

Thus certain derivatives of ψ must decay. A condition of this type appears in [2, 5^0, p. 112]. All other conditions are the same as in Example 1.

LEMMA 3. *If $k_0 \in R$ is such that $(r-A)^2+n(1-1/q) > (1+n/q)(r-A)$ for all $r \in R$, $r \geqq k_0$ then there exists $\varepsilon > 0$ such that for all $r \geqq k_0$ there exists a continuous function ω_2 on $[0, T]$ satisfying*

$$|K_t(f(u)(s))|_{r+\varepsilon,q} \leqq \tilde{\kappa}(t)\omega_2(|u(s)|_{r,q})|u(s)|_{r+\varepsilon,q}$$

for all s, $t \in [0, T]$.

Proof. Lemmas 1 and 2 settle the case when $r \in Z$ and $\varepsilon = 1$. The proof of Lemma 3 is a modification of this argument. It suffices to consider two cases, the first case is when $r+\varepsilon \in Z$. The second case is when $[r+\varepsilon] \leqq r < r+\varepsilon$ ($[r+\varepsilon]$ is the integral part of $r+\varepsilon$).

Case (i). Suppose that $r+\varepsilon \in Z$. The difficulty is proving Lemma 1 arose when trying to satisfy (19). In this case $|\beta| = r+\varepsilon-A$ and S becomes

$$S = \left(\frac{n}{q}+|\nu_1|-r-\varepsilon\right)^+ + \sum_{i=2}^{|\alpha|}\left(\frac{n}{q}+|\nu_i|-r\right)^+.$$

As before, the case $r \geqq A+n/q$ presents little problem. If $r < A+n/q$ then the largest value of S occurs when $|\alpha| = r+\varepsilon-A$ and $|\nu_i| = A+1$. Thus

$$S \leqq \left(\frac{n}{q}+A+1-r-\varepsilon\right) + \sum_{i=2}^{r-A+\varepsilon}\left(\frac{n}{q}+A+1-r\right)$$

$$= -(r-A)^2+(r-A)\left(\frac{n}{q}+1-\varepsilon\right)+\varepsilon\frac{n}{q}.$$

The expression is bounded by $n/p = n-n/q$ if

$$\varepsilon(A+n/q-r) \leqq (r-A)^2+n-\frac{n}{q}-\left(\frac{n}{q}+1\right)(r-A).$$

Therefore an appropriate value of $\varepsilon > 0$ exists if

$$(r-A)^2+n-\frac{n}{q} > \left(\frac{n}{q}+1\right)(r-A). \tag{24}$$

Case (ii). Suppose that $[r+\varepsilon] \leqq r < r+\varepsilon$. Let $k = [r+\varepsilon] = [r]$ and $|\beta| = k-A$. Here it is appropriate to obtain estimates in terms of the Besov spaces $B^r_{p,q}(R^n)$. See [2], [9], [10]. Let ω_p represent the L^p modulus of continuity of $g \in L^p(R^n)$:

$$\omega_p(\delta, g) = \sup_{|h| \leqq \delta} |g(\cdot+h)-g(\cdot)|_p, \qquad p \geqq 1.$$

Then the norm of $B^r_{p,q}(R^n)$ is given by

$$|||g|||_{r,p,q} \equiv |g|_p + \left(\int_0^\infty \left(\delta^{-r+k} \sum_{|\alpha|=k} \omega_p(\delta, \partial^\alpha g) \right)^q \frac{d\delta}{\delta} \right)^{1/q}.$$

These spaces are close to the $L^p_r(R^n)$ spaces in the sense that for any $r \in R$, $1 < p < \infty$, $1 < q \leqq \infty$, $\varepsilon > 0$,

$$L^p_{r+\varepsilon}(R^n) \subset B^r_{p,q}(R^n) \subset L^p_{r-\varepsilon}(R^n),$$

and the inclusions are continuous.

Here $|\beta| = k - A$, and we will prove that

$$|||\partial^\beta f_1(u)(t)|||_{r+\varepsilon_1-\varepsilon-k,p,q} \leqq \omega_2(|u(t)|_{r,q}) |||u(t)|||_{r-\varepsilon_1,q,q},$$

where $\varepsilon_1 > 0$ and $k \leqq r + \varepsilon_1 < k+1$. If we can prove this inequality then the statement of the lemma will follow by replacing the Besov space norms with L^p_r norms and then applying the operator K_t.

The function $\partial^\beta(f_1(u)(t))$ can be written as a sum of products as in (15). To calculate the L^p modulus of continuity of the terms where $|\alpha| \leqq \gamma$ it is necessary to prove that

$$|\partial_t^b \partial_x^\sigma \partial_z^\alpha \psi_1(t, x, z_1) - \partial_t^b \partial_x^\sigma \partial_z^\alpha \psi_1(t, x, z_2)| \leqq C(|z_1| + |z_2|)^{\gamma - |\alpha| - 1} |z_1 - z_2| \quad \text{if } |\alpha| \leqq \gamma - 1. \tag{25}$$

This is an easy calculation. However, when $\gamma - 1 < |\alpha| < \gamma$ the extra condition (23) is needed to prove that

$$|\partial_t^b \partial_x^\sigma \partial_z^\alpha \psi_1(t, x, z_1) - \partial_t^b \partial_x^\sigma \partial_z^\alpha \psi_1(t, x, z_2)| \leqq C|z_1 - z_2|^{\gamma - |\alpha|}. \tag{26}$$

With these two inequalities it is not hard to calculate the Besov space norm of the terms with $|\alpha| \leqq \gamma$. The Besov norms for the terms with $|\alpha| > \gamma$ are straightforward. The only problem is to check the condition imposed on the parameter r.

Equation (18) is replaced by

$$\left|\left|\left| \prod_{i=1}^{|\alpha|} \partial^{\nu_i} u(t) \right|\right|\right|_{r+\varepsilon_1-k,p,q} \leqq \sum_{j=1}^{|\alpha|} |||\partial^{\nu_j} u(t)|||_{r+\varepsilon_1-k,p/a_j,q} \prod_{i \neq j} |\partial^{\nu_i} u(t)|_{p/a_i}$$
$$\leqq C|u(t)|_{r,q}^{|\alpha|-1} |u(t)|_{r+\varepsilon,q},$$

where $\varepsilon_1 < \varepsilon$. The sum over j comes from the use of the triangle inequality in calculating the modulus of continuity. The conditions analogous to (19) are

$$\frac{p}{n}\left(\frac{n}{q} + |\nu_i| + \varepsilon_1 - \varepsilon - k\right)^+ \leqq a_j \leqq \frac{p}{q} \quad \text{if } i = j$$

$$\frac{p}{n}\left(\frac{n}{q} + |\nu_i| - r\right)^+ \leqq a_i \leqq \frac{p}{q} \quad \text{if } i \neq j.$$

We must ensure that

$$\left(\frac{n}{q} + |\nu_j| + \varepsilon_1 - \varepsilon - k\right)^+ + \sum_{i \neq j} \left(\frac{n}{q} + |\nu_i| - r\right)^+ < \frac{n}{p}. \tag{27}$$

As in Lemma 1 the case $r \geqq A + n/q$ is easy. Suppose that $r < A + n/q$. This sum is largest when $|\alpha| = k - A$ and $|\nu_i| = A + 1$. Now (27) becomes

$$\left(\frac{n}{q} + A + 1 - r\right)(k - A) + (\varepsilon_1 - \varepsilon - k + r) < \frac{n}{p}$$

or

$$(r-A)^2+n-\frac{n}{q}>(r-A)\left(\frac{n}{q}+1\right)-(r-k)\left(A+\frac{n}{q}+2-r\right)-(\varepsilon-\varepsilon_1).$$

Since this condition is weaker than (24) we are done. Therefore in this example k_0 can be any real number such that $k_0 \geqq A$ and

$$(r-A)^2+\gamma\left(\frac{n}{q}\right)>\left(1+\frac{n}{q}\right)(r-A)$$

for all $r \geqq k_0$.

LEMMA 4. *Let* $r \in R$, $r \geqq k$ *then*

(i) $|K_t(f(u)(s))|_{r,q} \leqq \tilde{\kappa}(t)\omega(|u(s)|_{r,q})|u(s)|_{r,q}$

(ii) $|K_t(f(v_1)(s)-f(v_2)(s))|_{r,q} \leqq \tilde{\kappa}(t)\omega_1(s)|v_1(s)-v_2(s)|_{r,q}^d$

where $\omega_1(s)=\omega_2(|v_1(s)|+|v_2(s)|)$ *and* ω_2 *is a continuous function on* $[0, T]$.

Proof. The power d comes from (23) and (26). Statement (i) follows from the proof of Lemma 3 with $\varepsilon=0$. The proof of (ii) is virtually the same. □

Example 3. The fact that L^q to L^p estimates were used for K_t resulted in the restriction (11) being imposed on γ. The growth condition on γ can be relaxed for $m>1$ if we assume that

$$|\partial_t^b\partial_x^\sigma\partial_z^\alpha\psi(t, x, z)| \leqq C|z|^{\max(\gamma-|\alpha|,0)} \tag{28}$$

and $\partial_t^b\partial_x^\delta\psi(t, x, 0)=0$ for all (t, x, z) and (b, δ, α). This is not a serious restriction in higher dimensions, where $\gamma<2$. However, where γ is permitted to be large it requires that ψ has a certain number of zero derivatives at $z=0$. In this case $\tilde{\kappa}(t)=\kappa(t)$ and K_t is considered only as a mapping from L^p to L^q. The proof of the analogue of Lemma 1 is easier here since the splitting $f=f_0+f_1$ and the ϕ_j are not necessary. As in the proof of Lemma 1, γ must satisfy

$$\gamma \leqq 1+n(q-2)/(n-(k-A)q).$$

If the upper bound in (7) is used for q then the new range is

$$1 \leqq \gamma<1+(4m-2A)/(n-2m+A-2(k_0-A)) \quad \text{if } m>1. \tag{29}$$

For the Schrödinger equation (22) and (9) imply that

$$1 \leqq \gamma<1+4m/(n-2m-2n(k_0-A)) \quad \text{if } 2m+2(k_0-A)<n$$

and

$$1 \leqq \gamma<\infty \quad \text{if } 2m+2(k_0-A) \geqq n.$$

Example 4. In some cases the critical index k_0 can be lowered. Let $0 \leqq A_1 \leqq A$ and suppose that $f(u)$ is a function as in Example 1 except that $f(u)$ is a polynomial as a function of $\partial^\alpha u$ for all α, $|\alpha|>A_1$. The nonlinearities studied by Pecher [7] and Narazaki [5] have the form $g(u)(t)=\sum_{|\beta| \leqq A/2}\partial^\beta(g_\beta(\partial^\beta u(t)))$, where g_β is a smooth function on R such that

$$|g_\beta^{(j)}(r)| \leqq C(1+|r|)^{\max(\gamma-j,0)}, \quad r \in R, j \geqq 0$$

and $g(0)=0$. This is a special case where ψ is a polynomial in $\partial^\alpha u$ for $|\alpha|>A/2$ (A is even). In this case from (20)

$$S \leqq (k+1-A/2)\left(\frac{n}{q}+A/2+1-k\right)-1$$

and the restriction on k becomes

$$(k-A/2)^2+n-\frac{n}{q}>(k-A/2)\frac{n}{q}.$$

The condition for γ is $\gamma\leqq 1+n(q-2)/(n-(k-A/2)q)$. This can be relaxed as in (29) if

$$|g_\beta^{(j)}(r)|\leqq C|r|^{\max(\gamma-j,0)},\quad r\in R$$

Example 5. Like polynomials, nonlinearities involving convolution can be quite simple. For example, if $f(u)=(V*u^2)u$ where $V\in L^r$, $n/2\leqq r\leqq\infty$, then

$$|f(u)|_{k+1,q}\leqq C|u(t)|_{k,q}^2|u(t)|_{k+1,q},$$

$A=0$, $\gamma=3$, and we may take $k_0=0$. Here we use estimates of K_t as an operator from L^q to L^q. The fact that $k_0=0$ will imply by Theorem 3 that all solutions with this type of nonlinearity are classical solutions, assuming that V and the initial conditions are smooth.

The implications of all of these inequalities for $f(u)$ will be discussed in §§ 5 and 6.

4. Existence. Define $W_k^q(T)$ to be the space of continuous bounded functions from $[0, T]$ to $L_k^q(R^n)$ with norm

$$\|g\|=\sup\{|g(t)|_{k,q}: 0\leqq t\leqq T\}.$$

The integral equation corresponding to the nonlinear wave equation (1) is

$$u(t)=K_t^C(u_0)+K_t(u_1)+\int_0^t K_{t-s}(g(s)-f(u)(s))\,ds \tag{30}$$

The solution for the corresponding linear equation is

$$h(t)=K_t^C(u_0)+K_t(u_1)+\int_0^t K_{t-s}(g(s))\,ds.$$

Thus

$$u(t)=h(t)-\int_0^t K_{t-s}(f(u)(s))\,ds\equiv F(u)(t).$$

Rather than using u_0, u_1, g it will usually be convenient to express our regularity assumptions in terms of h, the solution of the linear equation. For the existence of solutions we use the standard procedure of setting up a contraction mapping in the space $W_k^q(T_1)$.

LEMMA 5. *For a nonlinear equation* (30) *with corresponding parameters* k_0, γ, q, $T\leqq\infty$, *if* $h\in W_k^q(T)$ *and* $k\geqq k_0$ *then there exists* T_1, $0<T_1<T$ *such that* (30) *has a solution* $u\in W_k^q(T_1)$.

Proof. Let $k=k_0$. By Lemma 4 or its equivalent

$$\left|\int_0^t K_{t-s}(f(u)(s))\,ds\right|_{k,q}\leqq C_1\int_0^t\tilde{\kappa}(t-s)\omega(|u(s)|_{k,q})|u(s)|_{k,q}\,ds.$$

We may assume that ω and ω_2 are increasing. Hence, using the norm of $W_k^q(T_1)$,

$$\|F(u)\|\leqq\|h\|+C_1\left(\int_0^{T_1}\tilde{\kappa}(s)\,ds\right)\omega(\|u\|)\|u\|. \tag{32}$$

This shows that F maps $W_k^q(T_1)$ to itself.

Similarly, by Lemma 4,

$$
\begin{aligned}
|F(v_1)(t)-F(v_2)(t)|_{k,q} &\leqq \left| \int_0^t K_{t-s}(f(v_1)(s)-f(v_2)(s))\, ds \right|_{k,q} \\
&\leqq C_2\left(\int_0^{T_1} \tilde{\kappa}(s)\, ds \right) \omega_2(\|v_1\|+\|v_2\|)\|v_1-v_2\|^d/2.
\end{aligned} \tag{33}
$$

Let $M > 2\|h\|$, and suppose that T_1 is so small that

$$
\left(C_1\omega(M) + C_2\omega_2(M)M^{d-1} \right) \int_0^{T_1} \tilde{\kappa}(s)\, ds < \frac{1}{2}.
$$

If $\|u\| \leqq M$, $\|v_1\| \leqq M$, $\|v_2\| \leqq M$ then (32) and (33) become

$$
\|F(u)\| \leqq M \quad \text{and} \quad \|F(v_1)-F(v_2)\| \leqq \|v_1-v_2\|/2.
$$

Thus F is a contraction of the ball of radius M in $W_k^q(T_1)$. Consequently F has a unique fixed point $u \in W_k^q(T_1)$.

5. Regularity. The proof of the regularity of u uses repeated applications of the following well-known inequality.

LEMMA 6 (Gronwall's inequality). *Let $T>0$. If $U \in C([0, T])$, $H \in C(R_+)$, and $\kappa_1 \in L^1_{loc}(R_+)$ are positive functions satisfying*

$$
U(t) \leqq H(t) + \int_0^t \kappa_1(t-s)U(s)\, ds
$$

for all $0 < t < T$, then there exists $\omega_2 \in C(R_+)$, depending only on κ_1, such that

$$
U(t) \leqq \omega_2(t) \max_{0 \leqq s \leqq t} H(s)
$$

for all $0 \leqq t < T$.

The proof is a straightforward induction on T.

LEMMA 7. *Suppose that $f(u)$ is a nonlinearity from § 3 with parameters k_0, γ, q, and T. For any $k \geqq k_0$, if $h \in W_k^q(T)$ and $u \in W_{k_0}^q(T)$ is a solution of the integral equation* (30), *then $u \in W_k^q(T)$.*

Proof. From the integral equation and (10),

$$
\begin{aligned}
|u(t)|_{k+1,q} &\leqq |h(t)|_{k+1,q} + \int_0^t \tilde{\kappa}(t-s)\omega(|u(s)|_{k,q})|u(s)|_{k+1,q}\, ds \\
&\leqq |h(t)|_{k+1,q} + \tilde{\omega} \int_0^t \tilde{\kappa}(t-s)|u(s)|_{k+1,q}\, ds
\end{aligned} \tag{34}
$$

when $\tilde{\omega} = \sup\{\omega|u(s)|_{k,q}): 0 \leqq s \leqq T\}$. The proof will be completed by induction on k. The lemma is obviously true when $k = k_0$. Suppose it holds also for some $k \geqq k_0$. By Lemma 5 there exists $T_1 \leqq T$ such that $u \in W_{k+1}^q(T_1)$. Let T_1 be the largest such constant. Gronwall's inequality applied to (34) shows that

$$
|u(t)|_{k+1,q} \leqq \omega_2(t)\|h\|
$$

for $0 \leqq t \leqq T_1$, where ω_2 depends on $\tilde{\kappa}$ and $\tilde{\omega}$ but not on $|u(t)|_{k+1,q}$. If $T_1 < T$ then $\lim_{t \to T_1} |u(t)|_{k+1,q} < \infty$ and $u(t)$ can be extended past T_1 by a contraction argument as in Lemma 5. Since this contradicts the maximality of T_1 then $T_1 = T$. As t approaches T, $|u(t)|_{k+1,q}$ remains bounded. Consequently $u \in W_{k+1}^q(T)$. In the case of a nonlinearity from Example 2 the argument is the same except that we conclude that $u \in W_{k+\varepsilon}^q(T)$. The result is the same in either case.

LEMMA 8. *If u is a solution of the integral equation* (30), $u \in W_k^q(T)$ *and* $h \in W_k^2(T)$, $k > A + n/q$, $k \geqq k_0$, *then* $u \in W_k^2(T)$.

Proof. Since by Sobolev's inequality $L_k^q \subset L_A^\infty$ then

$$|f(u)(t)|_2 \leqq C(1+|u(t)|_{A,\infty})^{\gamma-1}|u(t)|_{A,2}.$$

Because $A \leqq m$ and K_t maps L^2 to L_m^2 then an application of Gronwall's inequality to the integral equation shows that $u \in W_A^2(T_1)$.

Let β be a multi-index with $|\beta| = k - A + 1$. Since $u(t) \in L_A^q \subset L_A^\infty$ then as in the proof of Lemma 1

$$|\partial^\beta (f(u)(t))|_2 \leqq C\omega(|u(t)|_{k,q}) \sum_{\alpha,\nu} \left| \prod_{i=1}^{|\alpha|} \partial^{\nu_i} u(t) \right|_2. \tag{35}$$

To apply Hölder's and Sobolev's inequalities as in Lemma 1 it is necessary to find constants $a_i \in [0, 1]$ such that $\sum a_i = 1$ and

$$\begin{aligned} \left(\frac{1}{2} - \frac{k+1-|\nu_i|}{n}\right)^+ &\leqq \frac{a_i}{2} \leqq \frac{1}{2} \quad i = 1 \\ \left(\frac{1}{q} - \frac{k-|\nu_i|}{n}\right)^+ &\leqq \frac{a_i}{2} \leqq \frac{1}{q} \quad i \geqq 2. \end{aligned} \tag{36}$$

The sum of the right-hand sides is $\frac{1}{2} + (|\alpha| - 1)/q \geqq \frac{1}{2}$. Note that $|\nu_i| \leqq |\beta| + A = k + 1$ and $|\nu_i| \leqq k$ for $i \geqq 2$. Since $k > A + n/q$ then

$$\frac{1}{q} - \frac{k-|\nu_i|}{n} \leqq \frac{1}{n}(|\nu_i| - A).$$

This means that the sum of the left-hand sides of (36) is largest when $|\alpha| = 1$ and is $\leqq \frac{1}{2}$. Consequently an appropriate collection on $\{a_i\}$ does exist and

$$|\partial^\beta (f(u)(t))|_2 \leqq C\omega(|u(t)|_{k,q})|u(t)|_{k+1,2}.$$

hence $|f(u)(t)|_{k+1-A,2}$ is bounded by the same expression. Now an application of Gronwall's inequality completes the proof that $u \in W_{k+1}^2(T_1)$.

LEMMA 9. *Let $k > n/2$. If $\phi \in L_k^2(R^n)$ then $K_t(\phi)$ is differentiable in t, $\partial(K_t(\phi))/\partial t = K_t^C(\phi) \in C(R_+^{n+1}) \cap W_k^2(\infty)$.*

If $\phi \in L_{k+m}^2(R^n)$ then $K_t^C(\phi)$ is differentiable in t, $\partial(K_t^C(\phi))/\partial t = -(aI + (-\Delta)^m)K_t(\phi) \in C(R_+^{n+1}) \cap W_k^2(\infty)$.

Proof. Let $\xi_* = \sqrt{a + |\xi|^{2m}}$ and $\phi_* = (I - \Delta)^{k/2}\phi$. The function ϕ_* is continuous by Sobolev's theorem. The Lebesgue dominated convergence theorem implies that

$$\Phi_s(\xi) = \hat{\phi}_*(\xi)\left\{\frac{\sin((t+s)\xi_*) - \sin(t\xi_*)}{s\xi_*} - \cos(t\xi_*)\right\}$$

converges in L^2 norm to zero as $s \to 0$. Since

$$\left|\frac{K_{t+s}(\phi) - K_t(\phi)}{s} - K_t^C(\phi)\right|_\infty \leqq \left|\frac{1}{s}(K_{t+s}(\phi_*) - K_t(\phi_*)) - K_t^C(\phi_*)\right|_2 = |\Phi_s|_2,$$

then the difference quotients converge uniformly to $K_t^C(\phi) = \partial\phi/\partial t$. The Plancherel theorem also implies that $K_t(\phi)$ and $K_t^C(\phi)$ are continuous functions of t with values in $C(R^n)$ whenever $\phi \in L_k^2$. The proof that $K_t^C(\phi)$ has a continuous derivative is similar. The lemma applies also to $K_t(\phi(s))$, where $\phi \in W_k^2(T)$.

Let $\tilde{W}_{k,J}^2(T)$ be the space of functions u such that $\partial_t^j u \in W_k^2(T)$ for all $0 \leqq j \leqq J$.

LEMMA 10. *Suppose that* $k_1 > (J-1)m + (J+1)A + n/2$, $J \leqq 1$, *and* u *and* h *are solutions of* (30) *and* (31), *respectively. If* $h \in \tilde{W}^2_{k_1,J}(T)$ *and* $u \in W^2_{k_1}(T)$ *then* $u \in \tilde{W}^2_{k,J}(T)$, *where* $k = k_1 - (J-1)m - JA$.

Proof. Analogous to (15),

$$\partial_t^j(f(u)(t)) = \sum_b \sum_\alpha (\partial_t^b \partial_z^\alpha \psi)(u)(t) \sum_\nu c_\nu \prod_{i=1}^{|\alpha|} \partial_t^{r_i} \partial_x^{\nu_i} u(t) \tag{37}$$

where $|\nu_i| \leqq A$, $|\tau_i| \leqq j$. If $u \in W^2_{k_2}$ and $k_2 > A + n/2$ then $\partial^\alpha u \in L^\infty$ for all $|\alpha| \leqq A$ and $(\partial_t^b \partial_z^\alpha \psi)(u)(t) \in L^\infty$. Also $\partial_t^{r_i} \partial_x^{\nu_i} u(t) \in L^\infty$ for all i. Therefore $u \in \tilde{W}^2_{k_2,j}(T)$ implies that $f(u) \in \tilde{W}^2_{k_2-A,j}(T)$. In particular, $f(u) \in \tilde{W}^2_{k_2-A,0}(T) = W^2_{k_2-A}(T)$.

Differentiating the integral equation (30) gives

$$\partial_t u = \partial_t h - \int_0^t K^C_{t-s}(f(u)(s))\, ds$$

$$\partial_t^2 u = \partial_t^2 h - f(u)(t) + \int_0^t (aI + (-\Delta)^m) K_{t-s}(f(u)(s))\, ds.$$

From the first equation $\partial_t u \in W^2_{k_1-A}(T)$, $u \in \tilde{W}^2_{k_1-A,1}(T)$, and $f(u) \in \tilde{W}^2_{k_1-2A,1}(T)$. Since the integrand is in $\tilde{W}^2_{k_1-m-2A}(T)$, the second equation gives $\partial_t^2 u \in W^2_{k_1-m-2A}(T)$, $u \in \tilde{W}^2_{k_1-m-2A,2}(T)$, and $f(u) \in \tilde{W}^2_{k_1-m-3A,2}(T)$. In general, $\partial_t^j u \in \tilde{W}^2_{k_1-(j-1)m-jA,2}(T)$. This proves the lemma.

THEOREM 3. *Let* $f(u)$ *be the nonlinearity described in* § 3 *with parameters* k_0, γ, q *and* T_1. *Let* $T \leqq T_1$, $k \geqq 1$, $k_1 > k(m+A) + n/2$, $k_1 \geqq k_0$. *Suppose that* $u_0 \in L^2_{k_1+km}(R^n)$, $u_1 \in L^2_{k_1+km-m}(R^n)$, $g \in \tilde{W}^2_{k_1-m,k}(T)$.

If $u \in W^q_{k_0}(T)$ *is a solution of the integral equation* (25) *then* $u \in \tilde{W}^2_{km,k}(T)$. *If* $k \geqq J$ *and* $km > J + n/2$ *then* $u \in C^J([0, T] \times R^n)$. *In particular, if* $J = 2$ *then* u *is a classical solution of the differential equation* (1).

Proof. Since $k_1 > n/2$, Lemma 9 proves that $K^C_t(u_0)$, $K_t(u_1)$ and $K_{t-s}(g(s))$ are in $\tilde{W}^2_{k_1,k}$. Therefore $h \in \tilde{W}^2_{k_1,k}(T)$. Lemmas 7 and 8 combine to show that $u \in W^q_{k_0}(T)$ implies $u \in W^q_{k_1}(T)$. Since $k_1 > k(m+A) + n/2$ then Lemma 10 proves that $u \in \tilde{W}^2_{km,k}(T)$.

This theorem applies also in the case where blow-up occurs: $T \leqq \infty$. If u has a critical amount of regularity k_0 then the smoothness of u_0, u_1, and g imply that u is smooth. Conditions that guarantee this critical amount of smoothness are discussed in the next section. It turns out that the conditions for solutions to be classical for $0 \leqq t \leqq T$ are similar to the conditions proved elsewhere for global classical solutions. Also, note from (11) that the power γ can increase as k_0 does.

The main difficulty for a specific equation is calculating the best values of k_0 such that

$$|K_t(f(u)(s))|_{k+\varepsilon,q} \leqq \tilde{\kappa}(t)\omega(|u(s)|_{k,q})|u(s)|_{k+\varepsilon,q}$$

for all $k \geqq k_0$ (the general form of (10)). For the nonlinearities of Example 1 satisfying (12), $k_0 - A$ is either zero or the largest root of the quadratic $x^2 - xn/q + n(1-2/q) = 0$ (Lemma 2). In Example 2 with the extra condition (23), $k_0 - A$ is either zero or the larger root of the quadratic $x^2 - (1+n/q)x + n(1-1/q) = 0$ (Lemma 3). In Example 5, $k_0 - A$ is zero. If $k_0 - A = 0$ then every solution in W^q_A with smooth initial conditions u_0, u_1, g will be smooth. If we take Example 2 with the upper bound given for q in (6),

$$k_0 = \frac{\{(n^2+n+2) + \sqrt{(n^2+n+2)^2 - 8n(n+3)(n+1)}\}}{n+1}, \quad \text{if } m = 1, \quad n \geqq 10.$$

As $n \to \infty$, $q \to 2$, and $k_0/n \to 1/2$. If $k_0 > n/2$ then Sobolev's theorem guarantees that $u \in L^\infty$. A similar earlier result by von Wahl [11, p. 269] proved that if $f(u)$ has polynomial growth, and $u \in W^2_{k_0}(T)$ where $k_0 \in Z$, $k_0 \geqq n/2$, then u is a classical solution of (1).

6. Classical solutions and energy estimates. Let $f(u)$ be a nonlinearity as in Example 2 with the extra condition (28). For simplicity assume that u_0, u_1, and g are sufficiently smooth. In this section we derive several simple consequences of Theorem 3.

If $m = 1$ then

$$1 \leqq \gamma \leqq 1 + 4n(1-A)/(n(n-1+2A) - 2(n+1)(k_0 - A)). \tag{38}$$

In this case $q = 2(n+1)/(n-1+2A)$ and k_0 can be calculated from Lemma 3; that is,

$$(k-A)^2 + n(1-1/q) > (1+n/q)(k-A) \quad \text{for all } k \geqq k_0.$$

This quadratic in $k_0 - A$ has no real roots if $(1+n/q)^2 < 4n(1-1/q)$ or

$$(n^2+n+2+2An)^2 < 8n(n+3-2A)(n+1). \tag{39}$$

in particular, if $A = 0$ then this is satisfied for $n < 10$. Theorem 3 implies that a solution of the integral equation (30) such that $u \in W^q_0(T)$ is a classical solution of (1) for $0 < t < T$ if $n < 10$. Similarly $u \in W^q_A(T)$ is a classical solution whenever n satisfies (39).

If $f(u)$ is a nonlinearity from Example 4 of the form

$$f(u)(t) = \sum_{|\beta| \leqq A/2} (-1)^{|\beta|} \partial^\beta (f_\beta(\partial^\beta u(t)))$$

where f_β is a smooth real-valued function on R, and $f_\beta(0) = 0$. We make the extra assumption that

$$\int_0^r f_\beta(s)\, ds \geqq 0 \quad r \in R.$$

For these nonlinearities the energy norm

$$E(t) = |\partial_t u(t)|^2_2 + |u(t)|^2_{m,2}$$

is bounded as a function of t for all t. This shows that a weak L^2 solution $u(t)$ exists for $0 \leqq t < \infty$. Suppose for simplicity that $m = 1$. It follows from Hölder's and Sobolev's inequalities that $u \in W^q_1(\infty)$ since

$$|\partial_i (f(u)(t))|_p \leqq |u(t)|^{\gamma-1}_{1,2} |\partial_i u(t)|_q$$

if $q \leqq \gamma + 1 \leqq 2n(q-2)/q(n-2) + 2$. The restriction on q leads to the following upper bound on γ:

$$\gamma < 1 + 4n(1-A)/(n-2)(n+1). \tag{40}$$

If $k_0 = 1$ then (38) becomes $1 \leqq \gamma \leqq 1 + 4n(1-A)/(n^2-3n-2-2A)$. This upper bound is slightly larger than $1 + 4/(n-3)$ when $A = 0$. The limit on γ is imposed therefore by the transition from L^2_1 to L^q_1 rather than by Theorem 3. By Theorem 3, $u(t)$ is a classical solution if $n < 10$ and (40) holds. The power (40) also appears in Lemma III.5 of [2], however, for lower dimensions, $n < 10$, Brenner and von Wahl are able to prove the existence of classical solutions for $\gamma < 1 + 4/(n-2)$.

If $m > 1$ then from (7) $q < 2n/(n-2m+A)$. The condition for k_0 is

$$(k-A)^2 + n(1-(n-2m+A)/2n) > (1+n(n-2m+A)/2n)(k-A) \quad \text{if } k \geqq k_0. \tag{41}$$

This quadratic has no real roots if

$$(n-2m+A+2)^2 < 8(n+2m-A).$$

Let $A=0$. From (29), $1 \leqq \gamma < 1+4m/(n-2m-2k_0)$. From Theorem 3 it follows that a solution u of (30) in $W_0^q(T)$ is a classical solution of (1) if $n<2m+2+4\sqrt{2m}$. Similarly a weak L^q solution ($u \in W_m^q(T)$) is a classical solution of (1) for $0<t<T$ if $n<4m+4m/(m-1)$ and a strong L^q solution ($u \in W_{2m}^q(T)$) is classical if $n<6m-m/(m-1/2)$.

These results appear quite weak when compared to the results in [7] and [6]. In these papers there is a positivity assumption to guarantee a finite energy norm. The difference is that in this case, in contrast to $m=1$, more essential use is made of the energy norm. See, for example, Lemma 9 in [6].

For the Schrödinger equation, $A=0$ and $\gamma < 1+4m/(n-2m-2k_0)$. The condition for the dimension is $(n-2m+2)^2<8(n+2m)$. Therefore every solution in $W_0^{\gamma+1}(T)$ is a classical solution for $0<t<T$ if $n<2m+2+4m\sqrt{2m}$. This agrees with the result of Tsutsumi and Hayashi, who assuming a finite energy norm, proved that for $m=1$, $n \leqq 9$, global classical solutions exist. The nonlinearities in this case typically are of the form $f_0(u)u$, where u is complex-valued and f_0 is a real-valued function on C.

Acknowledgment. I thank T. Narasaki for many interesting conversations on the regularity theorem and positivity conditions.

REFERENCES

[1] P. BRENNER, *On the existence of global smooth solutions of certain semi-linear hyperbolic equations*, Math. Z., 167 (1979), pp. 99-135.

[2] P. BRENNER AND H. VON WAHL, *Global classical solutions of nonlinear wave equations*, Math. Z., 176 (1981), pp. 87-121.

[3] L. HORMANDER, *Estimates for translation-invariant operators on L_p-spaces.* Acta Math., 104 (1960), pp. 93-145.

[4] K. JÖRGENS, *Das Anfangswertproblem im Grossen für eine Klasse nichtlineare Wellengleichungen*, Math. Z., 77 (1961), pp. 295-308.

[5] S. KLAINERMAN, *Global existence for nonlinear wave equations*, Comm. Pure and Appl. Math., 33 (1980), pp. 43-101.

[6] T. NARASAKI, *Global classical solutions of semilinear evolution equations.* Saitama Math. J., 4 (1986), pp. 11-33.

[7] H. PECHER, *L^p-Abschätzungen und Klassische Lösungen für nichtlineare Wellengleichungen*, I. Math. Z., 150 (1976), pp. 159-183; II. Manuscripta Math., 20 (1977), pp. 227-244.

[8] H. PECHER, *Existensatz für regurare Losungen semilineare Wellengleichungen*, Nachr. Akad. Wiss Göttingen, Math.-Phys. Kl., (1979), pp. 129-151.

[9] E. M. STEIN, Singular Integrals and Differentiability Properties of Functions, Princeton University Press, Princeton, NJ, 1970.

[10] M. TSUTSUMI AND N. HAYASHI, *Classical solutions of nonlinear Schrödinger equations in higher dimensions*, Math. Z., 177 (1981), pp. 217-234.

[11] W. VON WAHL, *Klassische Lösungen nichtlinearer Wellengleichungen im Großen*, Math. Z., 112 (1969), pp. 241-279.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 430-438, March 1989

011

INFINITE-ORDER DIFFERENTIAL EQUATIONS AND THE HEAT EQUATION*

NORMAN L. HILLS† AND JOHN M. IRWIN‡

Abstract. A method is given for solving certain infinite-order differential equations whose characteristic function is an entire function of order less than 1. The infinite-order operator is expressed as an infinite product of first-order operators and is then inverted by a sequence of integral operators. The method is a natural generalization of a finite-order method and can be computed numerically. The method is shown to converge and an error estimate is given. Applications to solutions of some heat equation problems are indicated.

Key words. infinite-order differential equation, entire functions, heat equation

AMS(MOS) subject classifications. primary 34A35, secondary 35K05

1. Introduction. It has seemed natural to many people to extend the idea of a linear ordinary differential equation with constant coefficients

$$\sum_{n=0}^{N} a_n f^{(n)}(t) = g(t)$$

to an equation with an infinite-order differential operator

$$\sum_{n=0}^{\infty} a_n f^{(n)}(t) = g(t). \tag{1}$$

In fact, during the first half of this century the subject received a great deal of attention. In 1936, Davis [3] published a book on the subject with a bibliography of over 40 pages. In more recent times infinite-order differential operators were used by Hirschman and Widder [4] to invert convolution transforms and by Widder [7] in his study of the heat equation. Significant work, frequently by Russian mathematicians, is still being done. We know of no efforts, however, to use modern numerical methods to solve practical problems related to equations of type (1).

Here we give a method for solving certain of these equations and a brief discussion showing how the method can be implemented numerically. We also give examples showing the connection of (1) to the heat equation. Many other types of infinite-order differential equations can be used for other problems, and we intend to treat some of them in later papers.

2. Basic assumptions and definitions. If we write (1) as $\sum_{n=0}^{\infty} a_n D^n f(t) = g(t)$, $(D = d/dt)$, then by analogy with the characteristic polynomial of a finite-order equation, the characteristic function is defined as $A(z) = \sum_{n=0}^{\infty} a_n z^n$. The infinite-order differential operator is then written as $A(D)$ and (1) becomes $A(D)f = g$.

To obtain a reasonable theory we must make certain assumptions about the functions $A(z)$, $f(t)$, and $g(t)$.

We will assume that $A(z)$ is an entire function of exponential type of order $\alpha < 1$. This means that for any $\varepsilon > 0$ there exists an N such that $|a_n| < n^{-n/(\alpha+\varepsilon)}$ for $n > N$. Except for polynomials, entire functions of order less than one always have an infinite number of zeros. We also assume that all zeros of $A(z)$ are real and negative. The

* Received by the editors October 7, 1987; accepted for publication (in revised form) June 9, 1988.
† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824.
‡ Department of Mathematics, Wayne State University, Detroit, Michigan 48202.

following examples: $\cosh(\sqrt{z})$, $z^{-1/2}\sinh(\sqrt{z})$, and the modified Bessel function $I_0(\sqrt{z})$, each of which is of order $\frac{1}{2}$, satisfy all our assumptions. (Our results can be immediately extended to the case where $A(z)$ has only a finite number of positive zeros.)

For $A(D)f$ to be defined, f must be infinitely differentiable and $\sum_{n=0}^{\infty} a_n f(t)^{(n)}$ must converge, but $f(t)$ need not be analytic.

To insure uniqueness of the solution of an Nth-order differential equation, N boundary conditions are prescribed. Here we prescribe the infinite number of conditions $f^{(n)}(t)\to 0$ for $t>0$, $t\to 0$. These conditions are required if $f(t)$ is to represent the temperature at an interior point of a region whose initial temperature is zero.

The function $g(t)$ is assumed to be continuous for $t\geqq 0$, and $g(t)=0$ if $t\leqq 0$. We do not assume $g(t)$ to be analytic or even differentiable. In our applications $g(t)$ will represent the temperature at the boundary of a region or the average over the boundary of the temperature. The behavior of $f(t)$ for negative t is of no concern.

In general our assumptions about $A(z)$ are more restrictive than those found in most of the classical literature, while the assumptions about $f(t)$ and $g(t)$ are considerably less restrictive.

3. Examples of applications to the heat equation. Consider the one-dimensional problem with variable end conditions:

$$\begin{aligned} &u_t(x,t)=u_{xx}(x,t), && -1<x<1, \quad 0<t,\\ &u(x,0)=0, && -1\leqq x\leqq 1,\\ &u(-1,t)=\phi_1(t), && 0\leqq t,\\ &u(1,t)=\phi_2(t), && 0\leqq t; \end{aligned}$$

$u(x,t)$ is required to be continuous for $-1\leqq x\leqq 1$, $0\leqq t$. We assume $\phi_1(t)$ and $\phi_2(t)$ are continuous for $t\geqq 0$ and vanish for $t\leqq 0$.

Let $g(t)=\frac{1}{2}[\phi_1(t)+\phi_2(t)]$ and let $f(t)=u(0,t)$. We will show that $f(t)$ satisfies the equation $\cosh(\sqrt{D})f(t)=g(t)$ together with the condition $f^{(n)}(0+)=0$. The expression

$$u(x,t)=\sum_{n=0}^{\infty}\left\{\frac{x^{2n}}{(2n)!}f^{(n)}(t)+\frac{x^{2n+1}}{(2n+1)!}h^{(n)}(t)\right\} \tag{2}$$

represents a solution to the heat equation if $f(t)$ and $h(t)$ are in a suitable class of infinitely differentiable functions. See Widder [7, p. 43] and Cannon [2, p. 27]. It is seen from this expression that $f(t)=u(0,t)$ and that

$$g(t)=\tfrac{1}{2}[u(1,t)+u(-1,t)]=\sum_{n=0}^{\infty}\frac{D^n}{(2n)!}f(t)=\cosh(\sqrt{D})f(t)$$

so that $\cosh(\sqrt{D})f(t)=g(t)$. As for the condition $f^{(n)}(0+)=0$, the fact that $\phi_1(t)$ and $\phi_2(t)$ are both zero for $t\leqq 0$ implies that $u(x,t)$ can be defined to be zero for negative t and will still satisfy the heat equation. In fact the extended $u(x,t)$ can be taken as the solution to the following problem whose boundary data are continuous:

$$\begin{aligned} &u_t(x,t)=u_{xx}(x,t), && -1<x<1, \quad -1<t,\\ &u(x,-1)=0, && -1\leqq x\leqq 1,\\ &u(-1,t)=\phi_1(t) && \text{if } 0\leqq t, \qquad u(-1,t)=0 \quad \text{if } t<0,\\ &u(1,t)=\phi_2(t) && \text{if } 0\leqq t, \qquad u(-1,t)=0 \quad \text{if } t<0. \end{aligned}$$

Since every solution to the heat equation is analytic in x and infinitely differentiable in t (Widder [7, p. 84]) we see that

$$f^{(n)}(0+) = \lim_{t\to 0+} \frac{\partial^n}{\partial t^n} u(0, t) = \frac{\partial^{2n}}{\partial x^{2n}} u(x, t)\Big|_{x=0} = 0.$$

In a similar manner we can show that $u_x(0, t) = h(t)$ satisfies $A(D)h(t) = \frac{1}{2}[\phi_1(t) - \phi_2(t)]$, where now $A(D) = \sinh(\sqrt{D})/\sqrt{D}$. It is also true that $h^{(n)}(0+) = 0$ for $n = 0, 1, 2, \cdots$.

If $f(t)$ is the temperature at the center of a long cylinder of radius 1 and if the temperature in the cylinder is zero at $t = 0$, then $f(t)$ will satisfy $I_0(\sqrt{D})f(t) = g(t)$, where $g(t)$ is the average temperature on the surface, $r = 1$, of the cylinder.

All these examples can be derived, at least formally, by using the Laplace transform to obtain an equation, $A(s)F(s) = G(s)$, where F and G are transforms of our f and g. The solution is then the inverse transform of $G(s)/A(s)$. The difficulties we encounter are that the transform of $g(t)$ and the inverse transform of G/A may be very difficult to find or may not even exist. If numerical methods are used, our method may be preferable since finding $f(t)$ may require less effort than the computation of $G(s)$ alone.

4. Infinite products and a method of solution. Entire functions of order less than one always have an infinite product expansion

$$A(z) = A(0) \prod_{n=0}^{\infty} \left(1 - \frac{z}{\lambda_n}\right) \tag{3}$$

where the λ_n are the zeros of $A(z)$. Moreover,

$$\frac{A'(z)}{A(z)} = \frac{d}{dz} \log A(z) = \sum_{n=0}^{\infty} (z - \lambda_n)^{-1}$$

so that

$$\frac{A'(0)}{A(0)} = -\sum_{n=0}^{\infty} \lambda_n^{-1}. \tag{4}$$

For functions of order α less than one it is also true that $\sum |\lambda_n|^{-\beta}$ converges for any $\beta > \alpha$. This will be used in the proof of Lemma 5.

For a discussion of all these facts see Chapters 1 and 2 of Boas [1]. Section 12.10 of [1] in which Boas discusses infinite-order differential equations with analytic $f(t)$ and $g(t)$ may also be of interest.

The operator $A(D)$ is now written as a product of operators:

$$A(D) = A(0) \prod_{n=0}^{\infty} \left(1 - \frac{D}{\lambda_n}\right). \tag{5}$$

The use of the infinite product rather than the infinite sum seems to be due to Ritt [6]. It has since been used by many others, e.g., Korobeinik [5], whose methods are related but differ considerably from those discussed here.

Although Ritt has made the same assumptions about the characteristic function $A(z)$ as we have, his objectives differ from ours. One result, however, his Theorem V, is important to us. He introduces the notation A and (A) as follows:

$$Af = \lim_{N\to\infty} A(0) \prod_{n=0}^{N} \left(1 - \frac{D}{\lambda_n}\right) f(t), \qquad (A)f = \lim_{N\to\infty} \sum_{n=0}^{N} a_n D^n f(t).$$

Theorem V states that $(A)f = Af$ if f is analytic. His proof, however, does not really require the analyticity of f. Infinite differentiability and the convergence of $(A)f$ will suffice. This more general result will be used in the proof of Theorem 2.

Equation (5) displays $A(D)$ as an infinite product of first-order operators. We denote them as $K_n = (1 - D/\lambda_n)$. To find the inverse of K_n we solve $K_n f(t) = f(t) - (1/\lambda_n) f'(t) = g(t)$ subject to the condition $f(0) = 0$. The well-known elementary solution is

$$f(t) = -\lambda_n e^{t\lambda_n} \int_0^t e^{-s\lambda_n} g(s)\, ds.$$

We define the operators L_n by

$$L_n g(t) = -\lambda_n e^{t\lambda_n} \int_0^t e^{-s\lambda_n} g(s)\, ds. \tag{6}$$

The L_n are inverses of the K_n in the sense that for any continuous g, $K_n L_n g = g$ and for any differentiable f satisfying $f(0) = 0$, $L_n K_n f = f$. We solve $A(D)f = g$ by the sequence of inverse operators, i.e., the solution to (1) is to be given by $f = \prod_{n=0}^{\infty} L_n g / A(0)$. In Theorem 1 we will prove the convergence of this infinite product of operators and in Theorem 2 we show that it indeed yields a solution to (1).

To construct a practical algorithm we let $f_0(t) = g(t)/A(0)$ and then define $f_{n+1}(t)$ recursively by $f_{n+1}(t) = L_n f_n(t)$. A numerical method for computing $L_n f_n(t)$ is given at the end of the paper.

THEOREM 1. *If $A(z)$ satisfies the above conditions, then for any continuous $g(t)$ the sequence $\{f_n\}$ defined by*

$$f_0(t) = \frac{g(t)}{A(0)}, \qquad f_{n+1}(t) = -\lambda_n e^{t\lambda_n} \int_0^t e^{-s\lambda_n} f_n(s)\, ds$$

converges uniformly in any bounded interval $0 \leqq t \leqq T$. We have the explicit error estimate

$$\|f - f_n\| < 2|\lambda_0| \, \|g\| A(0)^{-2} \left[A'(0) - A(0) \sum_{k=0}^{n-1} |\lambda_k|^{-1} \right].$$

Remarks. We use the sup norm: $\|f\| = \sup |f(t)|$ for $0 \leqq t \leqq T$ so that we can exploit the completeness of $C[0, T]$ and the fact that convergence in this norm is uniform convergence.

We will repeatedly use the formula

$$f_{n+1}(t) = -\lambda_n e^{t\lambda_n} \int_0^t e^{-s\lambda_n} f_n(s)\, ds \tag{7}$$

for the solution to the initial value problem

$$f_{n+1}(t) - (1/\lambda_n) f'_{n+1}(t) = f_n(t), \qquad f_{n+1}(0) = 0. \tag{8}$$

It is convenient to rearrange the first part of (8) as $f_{n+1}(t) - f_n(t) = (1/\lambda_n) f'_{n+1}(t)$ and take norms to get

$$\|f_{n+1} - f_n\| = |1/\lambda_n| \, \|f'_{n+1}\|. \tag{9}$$

The operator $L_n(g)$ in (6) may be written as $\int_0^t -\lambda_n e^{(t-s)\lambda_n} g(s)\, ds$. Define

$$\delta_n(t) = \begin{cases} -\lambda_n e^{t\lambda_n} & \text{if } t \geqq 0, \\ 0 & \text{if } t < 0, \end{cases} \tag{10}$$

so that

$$L_n(g) = \int_0^T \delta_n(t - s) g(s)\, ds. \tag{11}$$

Before proving Theorem 1 we prove two lemmas.

LEMMA 1. $\|f_{n+1}\| < \|f_0\|$ *for all* $n \geqq 0$.

Proof. Using (11) and taking norms, we have

$$\|f_{n+1}\| = \|L_n(f)\| = \sup \left| \int_0^T \delta_n(t-s) f_n(s)\, ds \right|$$

$$\leqq \|f_n\| \sup \left| \int_0^T \delta_n(t-s)\, ds \right|.$$

From (10) we compute $\int_0^T \delta_n(t-s)\, ds = 1 - e^{t\lambda_n}$ so that $\|f_{n+1}\| < \|f_n\| \sup |1 - e^{t\lambda_n}|$. Since $\lambda_n < 0$ and $t \geqq 0$, $|1 - e^{t\lambda_n}| < 1$. Hence $\|f_{n+1}\| < \|f_n\|$. Likewise $\|f_n\| < \|f_{n-1}\|$ etc., so that for all $n \geqq 0$, $\|f_{n+1}\| < \|f_0\|$. □

LEMMA 2. $\|f'_{n+1}\| < \|f'_1\|$ *for all* $n \geqq 1$.

Proof. We assume that $n \geqq 1$ and show that f'_{n+1} is obtained from f'_n in exactly the same way that f_{n+1} is obtained from f_n. Lemma 1 will then imply Lemma 2. We must assume $n \geqq 1$ since we did not assume $f_0(t) = g(t)/A(0)$ to be differentiable. But for $n \geqq 1$, $f_n(t)$ is differentiable and $f_n(0) = 0$, as can be seen from (7).

Let $t = 0$ in (8) and it follows that $f'_{n+1}(0) = 0$. We differentiate (8), and the result, together with $f'_{n+1}(0) = 0$, gives

$$f'_{n+1}(t) - (1/\lambda_n)(f'_{n+1}(t))' = f'_n(t), \quad f'_{n+1}(0) = 0, \quad (n \geqq 1). \tag{12}$$

We see that (12) is the same as (8) with f_n and f_{n+1} replaced by f'_n and f'_{n+1}, and Lemma 2 follows. □

Proof of Theorem 1. From (9), $\|f_{n+1} - f_n\| = |1/\lambda_n| \|f'_{n+1}\|$, and Lemma 2, $\|f'_{n+1}\| < \|f'_1\|$, we have

$$\|f_{n+1} - f_n\| < |1/\lambda_n| \|f'_1\|. \tag{13}$$

If $n < m$ are any two positive integers then $(f_n - f_m) = (f_n - f_{n+1}) + (f_{n+1} - f_{n+2}) + \cdots + (f_{m-1} - f_m)$. Taking norms, we have $\|f_n - f_m\| \leqq \|f_n - f_{n+1}\| + \|f_{n+1} - f_{n+2}\| + \cdots + \|f_{m-1} - f_m\|$. Using (13), we obtain $\|f_m - f_n\| < [|\lambda_n|^{-1} + |\lambda_{n+1}|^{-1} + \cdots + |\lambda_{m-1}|^{-1}] \|f'_1\|$. Hence

$$\|f_m - f_n\| < \|f'_1\| \sum_{k=n}^{m-1} |\lambda_k|^{-1}. \tag{14}$$

The right-hand side of (14) can be made arbitrarily small by making n sufficiently large. Therefore $\{f_n\}$ is a Cauchy sequence and $\lim_{m\to\infty} f_m = f$ exists.

To derive the error estimate we take the limit of (14) as $m \to \infty$ to obtain $\|f - f_n\| < \|f'_1\| \sum_{k=n}^{\infty} |\lambda_k|^{-1}$. Using (4), we have $\sum_{k=n}^{\infty} |\lambda_k|^{-1} = A'(0)/A(0) - \sum_{k=0}^{n-1} |\lambda_n|^{-1}$. From (9) with $n = 0$, we have $\|f'_1\| = |\lambda_0| \|f_1 - f_0\|$. But $\|f_1 - f_0\| < \|f_1\| + \|f_0\| < 2\|f_0\|$ so that $\|f'_1\| < 2|\lambda_0| \|f_0\|$. Since $f_0 = g/A(0)$ we have the final result:

$$\|f - f_n\| < 2|\lambda_0| \|g\| A(0)^{-2} \left[A'(0) - A(0) \sum_{k=0}^{n-1} |\lambda_k|^{-1} \right]. \tag{15}$$

Before proving Theorem 2 we need five more lemmas. The reader may wish to look at the proof of Theorem 2 before reading the proofs of these lemmas.

In the lemmas that follow $F(t)$ is any continuous function such that $F(0) = 0$ if $t < 0$.

LEMMA 3. *For any* $\varepsilon \geqq 0$, *there is an* ξ *with* $0 \leqq \xi \leqq \varepsilon$ *and a function* $r(t)$ *such that* $L_n F(t) = F(t - \xi) + r(t)$. *Moreover,* $F(t - \xi)$ *and* $r(t)$ *are continuous and* $\|r(t)\| \leqq 2e^{\varepsilon\lambda_n} \|F(t)\|$.

Proof. Since $\delta_n(t-s)=0$ if $s>t$, and $F(s)=0$ if $s\leqq 0$ we have

$$L_nF(t)=\int_0^T \delta_n(t-s)F(s)\,ds$$

$$=\int_{-\infty}^{t-\varepsilon}\delta_n(t-s)F(s)\,ds+\int_{t-\varepsilon}^{t}\delta_n(t-s)F(s)\,ds.$$

Because $\delta_n(t)$ is nonnegative, the generalized mean value theorem for integrals implies

$$\int_{-\infty}^{t-\varepsilon}\delta_n(t-s)F(s)\,ds=F(t-\zeta)\int_{-\infty}^{t-\varepsilon}\delta_n(t-s)\,ds \quad (\text{where } \varepsilon\leqq\zeta), \text{ and}$$

$$\int_{t-\varepsilon}^{t}\delta_n(t-s)F(s)\,ds=F(t-\xi)\int_{t-\varepsilon}^{t}\delta_n(t-s)\,ds \quad (\text{where } 0\leqq\xi\leqq\varepsilon).$$

From (10) we compute $\int_{-\infty}^{t-\varepsilon}\delta_n(t-s)\,ds=e^{\varepsilon\lambda}$ and $\int_{t-\varepsilon}^{t}\delta_n(t-s)\,ds=1-e^{\varepsilon\lambda}$ (where we write λ in place of λ_n to simplify notation). Even though ξ and ζ may be discontinuous functions of t, the functions $F(t-\xi)$ and $F(t-\zeta)$ are continuous since

$$F(t-\xi)=(1-e^{\varepsilon\lambda})^{-1}\int_{t-\varepsilon}^{t}\delta_n(t-s)F(s)\,ds, \quad \text{and}$$

$$F(t-\zeta)=e^{-\varepsilon\lambda}\int_0^{t-\varepsilon}\delta_n(t-s)F(s)\,ds.$$

We now have $L_nF(t)=[1-e^{\varepsilon\lambda}]F(t-\xi)+e^{\varepsilon\lambda}F(t-\zeta)$ or $L_nF(t)=F(t-\xi)-e^{\varepsilon\lambda}F(t-\xi)+e^{\varepsilon\lambda}F(t-\zeta)=F(t-\xi)+r(t)$, where $r(t)=-e^{\varepsilon\lambda}F(t-\xi)+e^{\varepsilon\lambda}F(t-\zeta)=e^{\varepsilon\lambda}[F(t-\zeta)-F(t-\xi)]$. Hence $\|r(t)\|\leqq 2e^{\varepsilon\lambda}\|F\|$. □

We return to subscripts and write ε_k and ξ_k in the next two lemmas, which deal with sequences of the operators L_k.

LEMMA 4. *If for each λ_k we choose an ε_k, then for each k there is an ξ_k with $0\leqq\xi_k\leqq\varepsilon_k$ and a function $R(t)$ such that $\prod_{k=n}^{m}L_k(F)=F(t-\chi)+R(t)$, where $\chi=\sum_{k=n}^{m}\xi_k$. Moreover, $F(t-\chi)$ and $R(t)$ are continuous and $\|R\|\leqq[\prod_{k=n}^{m}(1+2e^{\varepsilon_k\lambda_k})-1]\|F\|$.*

Proof. Use induction on the number, $m-n+1$, of factors in the product. If $m=n$, there is only a single $L_k=L_n$, and Lemma 3 gives the result. Assume the lemma is true for $m-n$ factors

$$\prod_{k=n}^{m-1}L_k(F)=F(t-\chi_{m-1})+R_{m-1},$$

where $\chi_{m-1}=\sum_{k=n}^{m-1}\xi_k$ and

$$\|R_{m-1}\|\leqq\left[\prod_{k=n}^{m-1}(1+2e^{\varepsilon_k\lambda_k})-1\right]\|F\|.$$

Applying L_m and using Lemma 3, we have

$$\prod_{k=n}^{m}L_kF=L_m\left[\prod_{k=n}^{m-1}L_k(F)\right]=L_m[F(t-\chi_{m-1})+R_{m-1}]$$

$$=F(t-\chi)+R_{m-1}(t-\xi_m)+r(t) \quad \text{where } \|r\|\leqq 2e^{\varepsilon_m\lambda_m}\|F+R_{m-1}\|.$$

Let $R = R_{m-1} + r$. Then R is continuous since r is continuous by Lemma 3 and R_{m-1} is continuous by the induction hypothesis. The function $F(t-\chi)$ is continuous since $F(t-\chi) = \prod_{k=n}^{m} L_k(F) - R(t)$. As for the norm of R, we have

$$\|R\| \leqq \|R_{m-1}\| + \|r\| \leqq \|R_{m-1}\| + 2e^{\varepsilon_m \lambda_m} \|F + R_{m-1}\|,$$

$$\|R\| \leqq (1 + 2e^{\varepsilon_m \lambda_m}) \|R_{m-1}\| + 2e^{\varepsilon_m \lambda_m} \|F\|.$$

Using the induction hypothesis on $\|R_{m-1}\|$, we have

$$\|R\| \leqq (1 + 2e^{\varepsilon_m \lambda_m}) \left[\prod_{k=n}^{m-1} (1 + 2e^{\varepsilon_k \lambda_k}) - 1 \right] \|F\| + 2e^{\varepsilon_m \lambda_m} \|F\|,$$

$$\|R\| \leqq \left[\prod_{k=n}^{m} (1 + 2e^{\varepsilon_k \lambda_k}) - 1 \right] \|F\|. \qquad \square$$

LEMMA 5. *There is a sequence of positive numbers ε_k such that $\sum_{k=n}^{\infty} \varepsilon_k$ and the product $\prod_{k=n}^{\infty} (1 + 2e^{\varepsilon_k \lambda_k})$ both converge.*

Proof. Since α, the order of $A(z)$, is less than one, there is a β such that $\alpha < \beta < 1$. Recall that the series $\sum_{n=0}^{\infty} |\lambda_n|^{-\beta}$ will converge if $\beta > \alpha$. We let $\varepsilon_k = |\lambda_k|^{-\beta}$ so that $\sum_{k=n}^{\infty} \varepsilon_k = \sum_{k=n}^{\infty} |\lambda_n|^{-\beta}$ converges.

The infinite product $\prod_{k=n}^{\infty} (1 + 2e^{\varepsilon_k \lambda_k})$ will converge if $\sum_{k=0}^{\infty} e^{\varepsilon_k \lambda_k} = \sum_{k=0}^{\infty} \exp(-|\lambda_k|^{(1-\beta)})$ converges. To show that this latter sum converges, let N be any fixed integer such that $N(1-\beta) > 1$. For any $x > 0$ we have $e^{-x} < N! x^{-N}$; hence $\exp(-|\lambda_k|^{(1-\beta)}) < N! |\lambda_k|^{-N(1-\beta)}$ and therefore $\sum_{k=n}^{\infty} \exp(-|\lambda_k|^{(1-\beta)}) < N! \sum_{k=n}^{\infty} |\lambda_k|^{-1} < \infty$. $\square$

LEMMA 6.

$$\lim_{n \to \infty} \prod_{k=n}^{\infty} L_k F = F.$$

Proof. Choose ε_k as in Lemma 5. The $F(t-\chi)$ and $R(t)$ of Lemma 4 are continuous and $\chi = \sum_{k=n}^{\infty} \xi_k$ converges since $0 \leqq \xi_k \leqq \varepsilon_k$. We can therefore take the limit as $m \to \infty$ in Lemma 4 to obtain $\prod_{k=n}^{\infty} L_k(F) = F(t-\chi) + R(t)$, where

$$\|R\| \leqq \left[\prod_{k=n}^{\infty} (1 + 2e^{\varepsilon_k \lambda_k}) - 1 \right] \|F\|.$$

Taking the limit as $n \to \infty$, we prove the lemma. $\square$

LEMMA 7.

$$\prod_{k=0}^{\infty} L_k F = \prod_{k=0}^{n-1} L_k \prod_{k=n}^{\infty} L_k F.$$

Proof. From the proof of Theorem 1 we know that $\prod_{k=n}^{\infty} L_k F$ converges uniformly so that $\prod_{k=0}^{n-1} L_k \prod_{k=n}^{\infty} L_k F = \prod_{k=n}^{\infty} L_k \prod_{k=0}^{n-1} L_k F = \prod_{k=0}^{\infty} L_k F$. $\square$

THEOREM 2. *If $f = A(0)^{-1} \prod_{k=0}^{\infty} L_k g$, then $A(D)f = g$, that is to say, the function constructed by Theorem 1 is a solution to (1).*

Proof. By Theorem V of [5], we have

$$A(D)f = \lim_{N \to \infty} \sum_{n=0}^{N} a_n D^n f(t) = \lim_{N \to \infty} A(0) \prod_{n=0}^{N} \left(1 - \frac{D}{\lambda_n}\right) f(t).$$

Introducing the definitions of f, K_n, L_n, and canceling $A(0)$, we have

$$A(D)f = \lim_{n \to \infty} A(0) \prod_{k=0}^{n} K_k \prod_{k=0}^{\infty} \frac{L_k g}{A(0)} = \lim_{n \to \infty} \prod_{k=0}^{n} K_k \prod_{k=0}^{\infty} L_k g.$$

Applying Lemma 7, we have,

$$A(D) = \lim_{n\to\infty} \prod_{k=0}^{n-1} K_k \prod_{k=0}^{n-1} L_k \prod_{k=n}^{\infty} L_k g.$$

Using the fact that the K_n and L_n are inverses, we have

$$A(D)f = \lim_{n\to\infty} \prod_{k=0}^{n-1} K_k \prod_{k=0}^{n-1} L_k \prod_{k=n}^{\infty} L_k g = \lim_{n\to\infty} \prod_{k=n}^{\infty} L_k g.$$

Applying Lemma 6, we have $A(D)f = \lim_{n\to\infty} \prod_{k=n}^{\infty} L_k g = g$. □

5. Numerical method for computing $L_n f(t)$. Our method necessitates computing $L_n f(t)$ for large values of $|\lambda_n|$. We will show how this can be done without computing the extremely large $e^{-s\lambda_n}$ or the extremely small $e^{t\lambda_n}$. To simplify notation we omit subscripts by writing $L(f) = F(t) = -\lambda\, e^{t\lambda} \int_0^t e^{-s\lambda} f(s)\, ds$. The integral is evaluated by standard interpolatory quadrature formulas with $e^{-s\lambda}$ as a weight function. Taking a stepsize of h, the two-point formula has the following form:

$$\int_t^{t+h} e^{-s\lambda} f(s)\, ds = \lambda^{-1}\, e^{(t+h)\lambda}[K_1 f(t) + K_2 f(t+h)].$$

Therefore

$$-\lambda\, e^{-(t+h)} \int_t^{t+h} e^{-s\lambda} f(s)\, ds = K_1 f(t) + K_2 f(t+h) \tag{17}$$

where

$$K_1 = \left[\frac{e^{h\lambda}-1}{-\lambda h} - e^{\lambda h}\right], \qquad K_2 = \left[1 - \frac{e^{h\lambda}-1}{-\lambda h}\right].$$

Note that the terms $\lambda^{-1}\, e^{(t+h)\lambda}$ and $\lambda\, e^{-(t+h)\lambda}$ cancel; thus K_1 and K_2 do not depend on the limits of integration but only on λ and the stepsize h. The same will happen if interpolatory quadrature formulas with more nodes are used, but we omit the details here.

A simple algorithm for computing $F(t) = L_n f$ can be based on (17).

Divide the interval $[0, T]$ into N subintervals each of length h. Let $\lambda = \lambda_n$ and compute K_1 and K_2. Let $K_3 = e^{\lambda h}$ and $F(0) = 0$. Then:

For $k = 1$ to N

Let $F[kh] = f[(k-1)h]K_1 + f[kh]K_2 + F[(k-1)h]K_3$

End of algorithm.

The formula for $F[kh]$ in this algorithm is the sum of the terms $f[(k-1)h]K_1 + f[kh]K_2$, which give $-\lambda\, e^{kh\lambda} \int_{k-1}^{kh} e^{-s\lambda} f(s)\, ds$ according to (15), and the term

$$F[(k-1)h]K_3 = -\lambda\, e^{(k-1)h\lambda} \int_0^{(k-1)h} e^{-s\lambda} f(s)\, ds\, e^{kh}.$$

These terms combine to give $-\lambda\, e^{kh\lambda} \int_0^{kh} e^{-s\lambda} f(s)\, ds$.

Note that as $\lambda \to -\infty$, $K_1 \to 0$ and $K_2 \to 1$ so that $F[kh] \to f[kh]$ agrees with Lemmas 3 and 6.

The algorithm has been tested on a variety of problems and found to be stable. When used on heat equation problems, whose solutions are known from other methods,

the error estimate has been found to be reliable. The speed of convergence can be substantially improved by standard "extrapolation to the limit" methods. We hope to report more fully on the numerical aspects of our method in a later paper.

REFERENCES

[1] R. P. BOAS JR., *Entire Functions*, Academic Press, New York, 1954.

[2] J. R. CANNON, *The One-Dimensional Heat Equation*, Encyclopedia of Mathematics Vol. 23, Addison-Wesley, Reading, MA, 1984.

[3] HAROLD THAYER DAVIS, *The Theory of Linear Operators from the Standpoint of Differential Equations of Infinite Order*, Principia Press of Illinois, Illinois, 1936.

[4] I. I. HIRSCHMAN AND D. V. WIDDER, *The Convolution Transform*, Princeton University Press, Princeton, NJ, 1955.

[5] Y. F. KOROBEINIK, *The existence of a solution of an equation of infinite order with a given type of growth*, Math. Notes, USSR, 13 (1973), pp. 406-410.

[6] J. F. RITT, *On a general class of linear homogeneous differential equations of infinite order with constant coefficients*, Trans. Amer. Math. Soc., 18 (1917), pp. 27-49.

[7] D. V. WIDDER, *The Heat Equation*, Academic Press, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 439-453, March 1989

012

ON THE BRANCHPOINT OPERATOR AND THE ANNIHILATION OF DIFFERINTEGRATIONS*

L. M. B. C. CAMPOS†

Abstract. A differintegration operator of rational, real or complex order ν, is defined in the literature [*Fractional Calculus and Applications*, Springer-Verlag, Berlin, New York, 1974; *Fractional Calculus and Integral Transforms of Generalized Functions*, Pitman, Boston, 1979; *IMA J. Appl. Math.*, 33 (1984), pp. 109-133; *Portugal. Math.*, 43 (1985), pp. 347-376], as a generalization of the ordinary nth derivative and nth primitive, which are the particular cases of order, respectively, a positive $\nu = +n$ or negative $\nu = -n$ integer $n \in \mathbb{N}$. In the present paper we introduce a branchpoint operator, which can annihilate the differintegration operator of any noninteger complex order, i.e., in all cases except ordinary derivatives and primitives. It is shown that branchpoint operator is distinct from, but related to, the rule of composition of differintegrations, and can arise if (a) the two systems of differintegration named after Liouville and Riemann are combined, or (b) if certain double-loop integrals are interchanged, in such a way that the inner path goes through a branchpoint. The differintegration operator has been used widely in connection with special functions [*Fractional Calculus*, Academic Press, New York, 1974; *Fractional Calculus*, 2 Vols., Descartes Press, Koryama, Japan 1984; *IMA J. Appl. Math.*, 36 (1986), pp. 191-206; *Mat. Vesnik*, 38 (1986), 375-390], and should not be confused with the branchpoint operator, which can lead to different formulas, as shown by the example of Hermite functions.

Key words. annihilation, differintegration, complex integration, cut-plane, branched functions, Hermite functions

AMS(MOS) subject classification. 30A99, 30C45, 30E20, 30E99, 33A65

1. Introduction.

The differintegration operator of complex order ν can be defined [16], [12] as a generalization of both the ordinary nth derivative and nth primitive, which correspond to the particular cases of order, respectively, $\nu = +n$, $-n$ a positive, negative integer $n \in \mathbb{N}$. There is one system of differintegration, or one type of derivative of complex order, for each set of branchcuts in the complex plane [1]; the two most important systems [21], [13] are the Liouville [11] operator D^ν / Dz^ν applying to analytic functions, and the Riemann [20] operator d^ν / dz^ν applying to functions with a branchpoint. The two systems are generally incompatible [2]; e.g., they lead to formulas for the differintegration of exponential (Liouville type) and power (Riemann type) that are inconsistent for nonintegral complex orders of derivation. The rules of Liouville and Riemann differintegration may be similar (e.g., for zero) or distinct (e.g., for a nonzero constant). Here we introduce a distinct, but related concept, namely, the branchpoint operator $\delta^\mu / \delta z^\mu$ of order μ.

The branchpoint operator arises if some "invalid" manipulations are performed with the differintegration operator, e.g.; (§ 2) if the operator is applied under integral sign at a branchpoint; (§ 3) if the order of integration in a double integral is interchanged so as to make one path go through a branchpoint. From this viewpoint some care should be taken when using the differintegration operator repeatedly because, when applied outside its conditions of validity, it may still give quite definite results that are no longer differintegrations, but rather the outcome of a distinct operator, the branchpoint operator. The branchpoint operator $\delta^\mu / \delta z^\mu$ can only be applied after the Liouville differintegration D^ν / Dz^ν, and the result is a Riemann differintegration $d^{\mu+\nu} / dz^{\mu+\nu}$, multiplied by a factor $A(\mu, \nu)$. Apart from the latter factor, the result looks like a

* Received by the editors October 21, 1986; accepted for publication (in revised form) May 6, 1988.

† Instituto Superior Técnico, 1096 Lisboa Codex, Portugal. The work of this author was supported by a research grant from J.N.I.C.T., and also by CAUTL, Instituto de Fisica-Matemática, INIC.

composition of derivatives, albeit of different types. Since the coefficient does not involve the independent z or dependent variables, the branchpoint operator retains the linear property of derivation operators. The coefficient $A(\mu, \nu)$ depends on the orders μ, ν of the operators, however, and can, in certain conditions (§ 3): (a) become equal to unity, in which case the branchpoint operator reduces to a differintegration; (b) vanish, in which case the branchpoint operator annihilates the differintegration to which it is applied. More precisely, the case (b) above shows that a branchpoint operator can be found such that it annihilates a Liouville differintegration operator of any complex order other than an integer (§ 4) so that the annihilation property fails to exist only for ordinary derivatives and primitives.

The differintegration operator provides a powerful method of study of the properties of special functions [16], [15], because the latter are specified by the differintegration of elementary functions [9], [8]. The application of rules of differintegration can be used to give convenient proofs of a number of properties of special functions, viz., both shorter proofs of known results [3] or new results [4] possibly harder to find by other methods. Since several of the proofs of properties of special functions involve repeated differintegrations, care must be taken not to confuse differintegration and branchpoint operators. This point is illustrated by the example of Hermite functions, which can be defined by the differintegration of the Gaussian function (§ 5); this definition agrees with the integral representation of Hermite functions of complex order and, in the case of positive integer order, reduces to the Hermite [7] polynomials. The branchpoint operator can be used to annihilate a Hermite function of nonintegral complex order, i.e., any Hermite function other than a polynomial. If the branchpoint operator is replaced by a differintegration, the result is no longer zero, but rather another Hermite function. The difference between branchpoint and differintegration operators is made clear by noting that they lead, when applied to Hermite functions, to distinct integral identities.

2. Relation between branchpoint and differintegration operators. Before we can define the branchpoint operator, we must recall the definition of differintegration operator, in the context [10], [14], [15], [1] of the theory of analytic functions.

DEFINITION 1. The derivative of complex order D^ν, of an analytic function $F(z)$, is defined by means of the generalized Cauchy integral:

$$\frac{D^\nu F}{Dz^\nu} \equiv \{\Gamma(1+\nu)/2\pi i\} \int_{\infty \exp\{i \arg(z)\}}^{(z+)} (\zeta - z)^{-\nu-1} F(\zeta)\, d\zeta, \tag{1}$$

along Fig. 1, a Hankel [6] loop going around $\zeta = z$ in the positive direction, and starting and ending at infinity, respectively, above and below the branchcut $C \equiv \{\zeta: |\zeta| > |z|, \arg(\zeta) = \arg(z)\}$. The integral is assumed uniformly convergent with regard to z in a region D and ν is not a negative integer.

Remark 1. The expression (1) with complex ν is designated "generalized Cauchy integral" because in the case where $\nu = n$ is a nonnegative integer $n \in \mathbb{N}_0$, the integrand

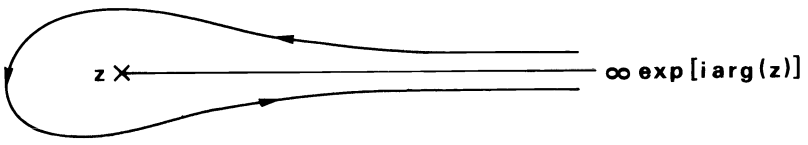


FIG. 1. *The differintegration of an analytic function involves a contour integral* (1) *along the Hankel* [6] *path, going counterclockwise around the branchpoint* $\zeta = z$, *and starting and ending at infinity in the* ζ*-plane, respectively, above and below the semi-infinite branchcut, joining* z *to infinity in the direction of* $\arg(z)$.

has a pole of order $n+1$, and the loop can be closed around it, so that we regain the original form of the Cauchy integral for the nth order derivative of an analytic function.

Remark 2. Formula (1) holds for all complex values of ν, other than negative integers $\nu=-n$ with $n\in\mathbb{N}$; this case is covered by the analytic continuation of (1) in the ν-plane for $\mathrm{Re}(\nu)<0$, by Weyl's (1917) formula:

$$\frac{D^{\nu}F}{dz^{\nu}}=\{\Gamma(-\nu)\}^{-1}\int_{\infty\exp\{i\arg(z)\}}^{z}(z-x)^{-\nu-1}F(x)\,dx, \tag{2}$$

which is proved elsewhere [1].

Remark 3. The convergence of the integrals (1) and (2) requires a restriction on the asymptotic behavior of the function $F(z)$ in a sector about the branchcut, e.g.,

$$\arg(\zeta)-\delta<\arg(z)<\arg(\zeta)+\delta: f(\zeta)\sim O(\zeta^{\mathrm{Re}(\nu)-\varepsilon}), \tag{3}$$

for some ε, $\delta>0$.

Remark 4. The case where $\nu=-n$ is a negative integer $n\in\mathbb{N}$, which was excluded from (1), and is included in (2), is the ordinary nth primitive.

The latter remark suggests that the preceding definition could be restated in terms of integration with complex order.

DEFINITION 2. The operator integration I^{ν} with complex order ν is defined by the derivative $D^{-\nu}$ of complex order $-\nu$.

Remark 5. The integration operator with complex order $I^{\nu}\equiv D^{-\nu}$ is defined by (1) with ν replaced by $-\nu$, for ν complex other than a positive integer; for $\mathrm{Re}(\nu)>0$, the integration operator I^{ν} is also specified by (2), with ν replaced by $-\nu$. In both cases an asymptotic condition is needed, e.g., (3) with $+\nu$ replaced by $-\nu$.

Remark 6. The similarity of Definitions 1 and 2, respectively, in terms of derivative and integral, has led to the adoption [21], [16] of the term "differintegration."

As an example, we consider the analytic function $F(z)=e^{az}$.

THEOREM 1. *The rule of differintegration of the exponential is*

$$\frac{D^{\nu}(e^{az})}{Dz^{\nu}}=a^{\nu}e^{az}. \tag{4}$$

Proof. Substitute $F(z)=e^{az}$ into (1) for ν complex other than a negative integer, and into (2) for $\mathrm{Re}\,(\nu)<0$. Evaluation of the integrals in terms of gamma functions leads [2] to (4).

Remark 7. Formula (4) was used by Liouville [11] to define a "derivation of nonintegral order."

Remark 8. Definition (1) can be designated, on historical precedent, as the "Liouville differintegration."

If function $F(z)$ has one branchpoint $z=a$, Definition (1) must be modified because now the branchcut is finite [9], [8], [1], [2], viz., Definition 3.

DEFINITION 3. The derivative of complex order ν of the function $(z-a)^{\mu}F(z)$, where $F(z)$ is analytic, and $(z-a)^{\mu}$ has a branchpoint at $\zeta=a$ with exponent μ, is defined by the generalized Cauchy integral:

$$\frac{d^{\nu}\{(z-a)^{\mu}F(z)\}}{dz^{\nu}}=\{\Gamma(1+\nu)/2\pi i\}\int_{a}^{(z+)}(\zeta-a)^{\mu}F(\zeta)(\zeta-z)^{-\nu-1}\,d\zeta, \tag{5}$$

along Fig. 2, a teardrop loop, going around $\zeta=z$ in the positive direction, and passing through $\zeta=a$. The integral is assumed to be uniformly convergent with regard to z in a region D, ν is complex other than a negative integer, and $\mathrm{Re}\,(\mu)>-1$.

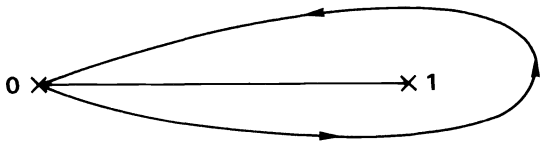


FIG. 2. *Teardrop loop in the t-plane, passing through the branchpoint* $t=0$ $(t=1)$ *and going in the positive direction around the other branchpoint* $t=1$ $(t=0)$, *so that the finite branchcut along the straight line* $(0,1)$ *lies in its interior.*

Remark 9. The definition can be extended [2] to all other values of μ, ν, by using integrals similar to (5), along: (i) the straight line joining $\zeta=a$ to $\zeta=z$; (ii) the teardrop loop, Fig. 3, passing through $\zeta=z$ and going around $\zeta=a$; and (iii) a Pochhammer [19] double-laced loop going around both $\zeta=a$ and $\zeta=z$, twice, in opposite directions.

THEOREM 2. *Definition* (5) *leads to the rule of derivation of the power*:

$$\frac{d^{\nu}(z^{\mu})}{dz^{\nu}}=\{\Gamma(1+\mu)/\Gamma(1+\mu-\nu)\}z^{\mu-\nu}, \tag{6}$$

with μ *not a negative integer.*

Proof. The proof is given in the literature (e.g., [2]).

Remark 10. Formula (6) was used as the definition of a "derivative of general order" by Riemann [20], so that (5) may be designated the Riemann system of differintegration.

The Liouville (1) and Riemann (5) differintegrations are the two simplest systems, since they involve only one branchcut in the complex plane, viz., respectively, infinite and finite. Other systems of differintegration involving more than one branchcut in the complex plane exist [1], but they are not needed to define the branchpoint operator. To introduce the latter, we consider a repeated Liouville (1) differintegration:

$$D^{\mu}\left(\frac{D^{\nu}F}{Dz^{\nu}}\right)\Big/Dz^{\mu}\equiv\{\Gamma(1+\mu)/2\pi i\}\int_{\infty\exp\{i\arg(z)\}}^{(z+)}(\zeta-z)^{-\mu-1}\left(\frac{D^{\nu}F}{D\zeta^{\nu}}\right)d\zeta, \tag{7}$$

where μ, ν are complex numbers other than negative integers. Substitution of (1) into (7) would lead to a double integral, the evaluation of which is deferred to § 3, since it leads to a less direct approach to the branchpoint operator. A simple and direct way of applying the differintegration of order μ to (1) would be to take the differintegration under integral sign, as a "Riemann differintegration":

$$\delta^{\mu}\left(\frac{D^{\nu}F}{Dz^{\nu}}\right)\Big/\delta z^{\mu}\equiv\{\Gamma(1+\nu)/2\pi i\}\int_{\infty\exp\{i\arg(z)\}}^{(z+)}F(\zeta)\frac{d^{\mu}\{(\zeta-z)^{-\nu-1}\}}{dz^{\mu}}d\zeta; \tag{8}$$

note that (8) is not equivalent to (7), because a Liouville differintegration D^{μ}/Dz^{μ}, outside the integral sign in (7), is replaced inside the integral sign in (8), by a Riemann differintegration d^{μ}/dz^{μ}, which is that appropriate (6) to the power $(\zeta-z)^{-\nu-1}$. Thus we are not calculating a repeated Liouville differintegration, as in (7), but rather (8) defines a new concept, which we call the branchpoint operator.

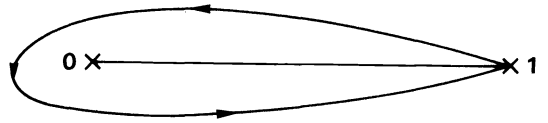


FIG. 3. *Mirror image of Fig.* 2.

DEFINITION 4. The branchpoint operator $\delta^\mu/\delta z^\mu$ of complex order μ, applied to the Liouville differintegration (1) of order ν, complex other than a negative integer, is defined by (8) applying a Riemann differintegration (5) under integral sign.

We have now to show that the branchpoint operator is unique, and indicate how it can be calculated.

THEOREM 3. *The branchpoint operator $\delta^\mu/\delta z^\mu$ of order μ, applied to a Liouville differintegration D^ν/Dz^ν of order ν, yields a Liouville differintegration of order $\mu+\nu$, to within a factor, in curly brackets, which depends only on μ, ν:*

$$\delta^\mu\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/\delta z^\mu = \{\Gamma(1+\nu)\Gamma(-\nu)/\Gamma(1+\mu+\nu)\Gamma(-\mu-\nu)e^{i\pi\mu}\}\frac{D^{\mu+\nu}F}{Dz^{\mu+\nu}}; \tag{9}$$

the formula holds for complex μ, ν, $\mu+\nu$, excluding integer ν and negative integer $\mu+\nu$.

Proof. Apply the Riemann rule of differintegration (6) to (8), to obtain

$$\begin{aligned}\delta^\mu\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/\delta z^\mu &= \{\Gamma(1+\nu)\Gamma(-\nu)/2\pi i\Gamma(-\nu-\mu)\}\cdot e^{-i\pi\mu}\\ &\quad\cdot\int_{\infty\exp\{i\arg(z)\}}^{(z+)} F(\zeta)(\zeta-z)^{-\mu-\nu-1}\,d\zeta,\end{aligned} \tag{10}$$

where we have used $\{d(-z)/dz\}^\mu = (-)^\mu = e^{-i\pi\mu}$, because the transformation $-z\to z$ along the Hankel loop is equivalent to a change of argument $-\pi$, i.e., argument π in the clockwise direction. Formula (6) applies in (8) if $-\nu-1$ is not a negative integer, i.e., ν is neither zero nor a positive integer; since ν, a negative integer, has already been excluded by (8) Definition 1, all integer values of ν are excluded in (10). The latter simplifies to (9), by using (1), for $\mu+\nu$, not a negative integer.

3. Derivation at and inversion of integrals through a branchpoint. Starting from the repeated Liouville differintegration (7), we may: (i) evaluate the double integral, leading to the rule of composition of differintegrations (Theorem 4); and (ii) interchange the integrals, so that one of the paths goes through a branchpoint, and the branchpoint operator is obtained instead (Theorem 5). Before we can indicate the deformation of path that leads (ii) to the branchpoint operator, we must revise the proof of the composition rule (i), to show where the change takes place.

THEOREM 4. *The Liouville differintegration operator satisfies the composition rule*:

$$D^\mu\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/Dz^\mu = \frac{D^{\mu+\nu}F}{Dz^{\mu+\nu}} = D^\nu\left(\frac{D^\mu F}{Dz^\mu}\right)\Big/Dz^\nu, \tag{11}$$

where μ, ν, $\mu+\nu$ are complex other than negative integers.

Proof. When we substitute (1) into (7), the repeated Liouville differintegration is given by

$$\begin{aligned}D^\mu\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/Dz^\mu &= -\{\Gamma(1+\nu)\Gamma(1+\mu)/4\pi^2\}\\ &\quad\cdot\int_{\infty\exp\{i\arg(\zeta)\}}^{(z_+)}(\zeta-z)^{-\mu-1}\\ &\quad\cdot\int_{\infty\exp\{i\arg(\zeta)\}}^{(\zeta+)}(\eta-\zeta)^{-\nu-1}F(\eta)\,d\eta\,d\zeta,\end{aligned} \tag{12}$$

with μ, ν complex other than negative integers. The $d\eta$ integration is performed (Fig. 4) along a Hankel loop around $\eta=\zeta$, and the $d\zeta$ integration along another Hankel

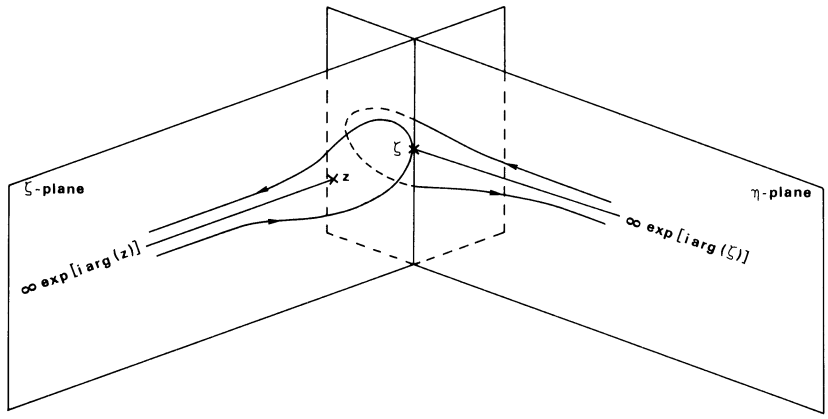


FIG. 4. *The double differintegration involves two Hankel paths in intersecting ζ- and η-planes, such that the branchpoint $\eta = \zeta$ in the η-plane lies on the path of integration in the ζ-plane and the branchpoint $\zeta = z$ in the ζ-plane lies on no path of integration, i.e., no path touches a branchpoint in its own plane.*

loop around $\zeta = z$, so that none of the loops touches a branchpoint in its own plane (although the Hankel loop in the ζ-plane touches the branchpoint $\zeta = \eta$ in the η-plane, which is not a problem). Since the integrals (12) are uniformly convergent, the order of integration can be reversed:

$$D^{\mu}\left(\frac{D^{\nu}F}{Dz^{\nu}}\right)\Big/ Dz^{\mu} = -\{\Gamma(1+\nu)\Gamma(1+\mu)/4\pi^2\}\int_{\infty \exp\{i \arg(z)\}}^{(z+)} F(\eta)\Phi(z,\eta)\, d\eta, \tag{13a}$$

where the function Φ is given by

$$\Phi(z,\eta) \equiv \int_{\infty \exp\{i \arg(\eta)\}}^{(\eta+)} (\zeta - z)^{-\mu-1}(\eta - \zeta)^{-\nu-1}\, d\zeta, \tag{13b}$$

which is an Eulerian integral of the second kind.

To evaluate (13b), we perform the change of variable $t = (\eta - z)/(\zeta - z)$, which maps the points $\zeta = \infty$, η to $t = 0$, 1, so that we obtain (Fig. 2) a teardrop loop passing through $t = 0$, and going counterclockwise around the branchpoint $t = 1$; the integral along this loop can be evaluated in terms [5] of beta functions:

$$(\eta - z)^{\mu+\nu+1}\Phi(z,\eta) = \int_0^{(1+)} t^{\mu+\nu}(t-1)^{-\nu-1}\, dt = -2i \sin(\pi\nu) B(\mu+\nu+1, -\nu); \tag{14a}$$

the integral (14a) converges only for $\mathrm{Re}\,(\mu+\nu) > -1$, but it can be extended, using a Pochhammer [19] double-laced loop, to [5] all nonintegral values of $\mu+\nu$. Thus (14a) holds for all complex values of $\mu+\nu$, except negative integers. The formula (14a) may be simplified using the relation between beta and gamma functions, and the symmetry properties of the latter:

$$\begin{aligned}\Phi(z,\eta) &= -2i \sin(\pi\nu)\Gamma(-\nu)\{\Gamma(1+\mu+\nu)/\Gamma(1+\mu)\}(\eta - z)^{-\mu-\nu-1} \\ &= 2\pi i\{\Gamma(1+\mu+\nu)/\Gamma(1+\mu)\Gamma(1+\nu)\}(\eta - z)^{-\mu-\nu-1}.\end{aligned} \tag{14b}$$

Substituting (14b) into (13a), we obtain

$$\begin{aligned}D^{\mu}\left(\frac{D^{\nu}F}{Dz^{\nu}}\right)\Big/ Dz^{\mu} &= \{\Gamma(1+\mu+\nu)/2\pi i\}\int_{\infty \exp\{i \arg(z)\}}^{(z+)} (\eta - z)^{-\mu-\nu-1}F(\eta)\, d\eta \\ &= \frac{D^{\mu+\nu}F}{Dz^{\mu+\nu}},\end{aligned} \tag{15}$$

where (1) was used, because $\mu+\nu$ is not a negative integer. The proof of (15), which coincides with (11) when the symmetry in $\mu+\nu$ is taken into account, thus holds for all complex μ, ν, $\mu+\nu$ other than negative integers. □

To obtain the branchpoint operator, in addition to the inversion of integrals (13a), (13b), we interlock the two paths of integration (Fig. 5), by replacing the inner Hankel path in (13b) with a teardrop loop.

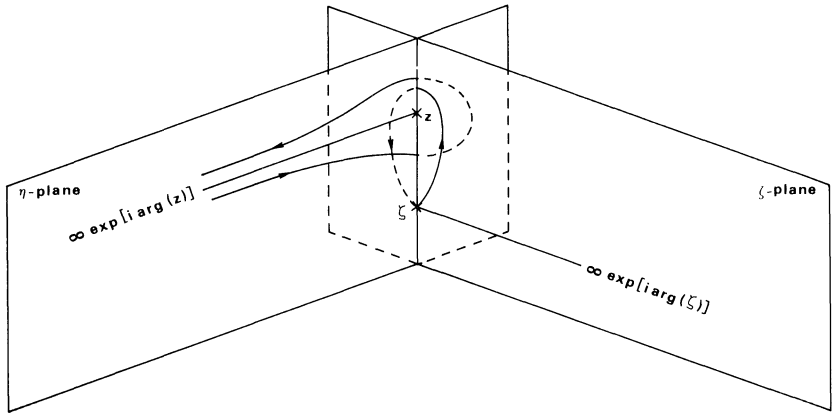


FIG. 5. *The deduction of the composition rule requires only the interchange of the η- and ζ-planes in Fig. 4, but if the first Hankel path is replaced with a teardrop loop, a distinct, branchpoint operator arises, because the two loops are interlocked and there is a finite branchcut along part of the intersection of the two planes.*

DEFINITION 5. The branchpoint operator $\delta^\mu/\delta z^\mu$ of order μ, applied to the Liouville differintegration D^ν/Dz^ν of order ν, not a negative integer, is defined as the composition of Liouville differintegrations (12) of orders μ, ν, in which the outer path of integration is unchanged (13a):

$$\text{(16a)}\quad \delta^\mu\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/\delta z^\mu \equiv -\{\Gamma(1+\nu)\Gamma(1+\mu)/4\pi^2\}\int_{\infty\exp\{i\arg(z)\}}^{(z+)} F(\eta)\psi(z,\eta)\,d\eta,$$

and the inner Hankel path (13b) in Fig. 4, is replaced with a teardrop loop in Fig. 5:

$$\text{(16b)}\qquad \psi(z,\eta)\equiv\int_{\zeta=\eta}^{(z+)}(\zeta-z)^{-\mu-1}(\eta-\zeta)^{-\nu-1}\,d\zeta,$$

which goes round $\zeta=z$ in the positive direction, and passes through the branchpoint $\zeta=\eta$.

Remark 11. The substitution of two Liouville differintegrations in (13a), (13b), with one Liouville and one Riemann differintegration in (16a), (16b), is analogous to the change from (8) to (9), to define the branchpoint operator.

Thus we may expect the present Definition 5 of branchpoint operator (16a), (16b), to be consistent with Definition 4, i.e., it leads to the same result (9), as was deduced from (7).

THEOREM 5 (Consistency of Definitions 4 and 5 of branchpoint operator). *The branchpoint operator, introduced* ((16a), (16b)) *by Definition* 5, *coincides with* (9), *which was introduced by Definition* 3:

$$\text{(17)}\qquad \delta^\mu\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/\delta z^\mu = e^{-i\pi\mu}\{\sin\{\pi(\mu+\nu)\}/\sin(\pi\nu)\}\frac{D^{\mu+\nu}F}{Dz^{\mu+\nu}}.$$

Remark 12. We will prove Theorem 5, using the same steps as for Theorem 4, to show that the only difference between the composition rule (11) and branchpoint operator (17) lies in the deformation of the inner path of integration from (13b) to (16b).

Thus we can start the proof by evaluating (16b).

Proof. To evaluate (16b), we may perform the change of variable $s = (\zeta - z)/(\eta - z)$, which is the inverse $s = 1/t$ of the transformation $t = (\eta - z)/(\zeta - z)$ used to evaluate (13b); the reason is that the inversion $s = 1/t$ maps a Hankel path in the t-plane to a teardrop loop in the s-plane. The transformation $s = (\zeta - z)/(\eta - z)$ maps the teardrop loop in (16b) into a rescaled loop, going round $\zeta = z$ or $s = 0$ in the positive direction and passing through $\zeta = \eta$ or $s = 1$; this teardrop loop, in Fig. 3, is the mirror image of that in Fig. 2, which was used in the evaluation of (14a). The latter is replaced, in the present case, with

$$(18a)\quad (\eta - z)^{\mu+\nu+1}\psi(z, \eta) = \int_1^{(0+)} s^{-\mu-1}(1-s)^{-\nu-1}\, ds = -2i \sin(\pi\mu)\, e^{-i\pi\mu} B(-\mu, -\nu),$$

where $B(\cdots, \cdots)$ denotes Euler's beta function, and the integral converges for $\mathrm{Re}\,(\nu) < 0$. By using the Pochhammer loop, it can be extended to all complex, noninteger values of ν. Thus the evaluation of (18a) holds for all ν, other than zero or positive integers. We had assumed earlier, in Definition 5, that ν is not a negative integer; thus, the derivation so far excludes all integer values of ν. The beta function is related [5] to the gamma function:

$$(18b)\quad \begin{aligned}\psi(z, \eta) &= -2i \sin(\pi\mu)\, e^{-i\pi\mu}\Gamma(-\mu)\{\Gamma(-\nu)/\Gamma(-\mu-\nu)\}(\eta - z)^{-\mu-\nu-1} \\ &= e^{-i\pi\mu}\{2\pi i\Gamma(1+\mu)\Gamma(-\nu)/\Gamma(-\mu-\nu)\}(\eta - z)^{-\mu-\nu-1}.\end{aligned}$$

Substituting (18b), which replaces (14b), into (16a), which replaces (13a), we obtain

$$\begin{aligned}\delta^\mu\left(\frac{D^\nu F}{Dz^\nu}\right) \Big/ \delta z^\mu &= \{\Gamma(1+\nu)\Gamma(-\nu)/\Gamma(1+\mu+\nu)\Gamma(-\mu-\nu)\}\, e^{-i\pi\mu} \\ &\quad \cdot \{\Gamma(1+\mu+\nu)/2\pi i\} \int_{\infty \exp\{i \arg(z)\}}^{(z+)} (\eta - z)^{-\mu-\nu-1} F(\eta)\, d\eta \qquad (19) \\ &= \{\sin\{\pi(\mu+\nu)\}/\sin(\pi\nu)\, e^{i\pi\mu}\} \frac{D^{\mu+\nu}F}{Dz^{\mu+\nu}},\end{aligned}$$

where we have used (1), assuming that $\mu + \nu$ is not a negative integer. The result (19) differs from the composition rule (15) in the term in curly brackets. The conditions of validity are the same on $\mu + \nu$, i.e., not a negative integer, but are distinct on: (i) μ, unrestricted in (19), not a negative integer in (15), because of the Liouville differintegration; (ii) ν, not a negative integer in (15) and not an integer in (19), i.e., the branchpoint operator does not apply to ordinary derivatives or primitives. □

4. Composition and annihilation of derivates of complex order. Definition 4 and Theorem 3 in § 2, or Definition 5 and Theorem 5 in § 3, give two distinct, but equivalent ((9) ≡ (17)) approaches to the branchpoint operator, which may be summarized in the following "anomalous" property A1.

PROPERTY 1. The repeated Liouville D^μ, D^ν differintegration (12), with orders μ, ν, if either (Definition 4) the outer differintegration is taken under integral sign as Riemann d^μ differintegration (8), or (Definition 5) in the inversion of integrals (13a), (13b) the inner Hankel path is deformed (16a), (16b) into a teardrop loop, then the

branchpoint operator δ^{μ} is obtained, which, when applied to the Liouville D^{ν} differintegration of order ν, differs from the compound Liouville $D^{\mu+\nu}$ differintegration of order $\mu+\nu$:

$$\delta^{\mu}\left(\frac{D^{\nu}F}{Dz^{\nu}}\right)\delta z^{\mu} = A(\mu,\nu)\frac{D^{\mu+\nu}F}{Dz^{\mu+\nu}}, \tag{20}$$

by a factor A, which does not involve the dependent $F(z)$ or independent z variables:

$$\begin{aligned} A(\mu,\nu) &\equiv e^{-i\pi\mu}\{\sin\{\pi(\mu+\nu)\}/\sin(\pi\nu)\} \\ &= \{1-\exp(-i\pi(\mu+\nu))\}/\{1-\exp(-i\pi\nu)\}, \end{aligned} \tag{21}$$

but depends only on the orders μ, ν, such that ν is not an integer, and $\mu+\nu$ not a negative integer.

Remark 13. The exclusion of $\mu+\nu$ a negative integer means that the result of a branchpoint operator applied to a Liouville differintegration cannot be an ordinary primitive.

Remark 14. The exclusion of ν an integer implies that the branchpoint operator applies only to Liouville differintegration other than ordinary derivatives or primitives for which there is no branchpoint in the integrand of (1).

Since the branchpoint operator δ^{μ_1} applied to a Liouville differintegration D^{ν} yields to within a factor not involving the dependent or independent variables, another Liouville differintegration $D^{\mu_1+\nu}$, a second branchpoint operator δ^{μ_2} may be applied, and so on, iteratively, up to δ^{μ_N} for any positive integer N. Concerning the factor (21), it can be decomposed into the following ratio:

$$A(\mu,\nu) = a(\mu+\nu)/a(\nu), \tag{22a}$$

$$a(\alpha) \equiv 1-\exp(-i2\pi\alpha), \tag{22b}$$

which satisfies the following recursive rule:

$$\begin{aligned} \prod_{k=1}^{N} A(\mu_k, \mu_{k-1}+\cdots+\mu_1+\nu) &= \prod_{k=1}^{N}\{a(\mu_k+\cdots+\mu_1+\nu)/a(\mu_{k-1}+\cdots+\mu_1+\nu)\} \\ &= a(\mu_N+\cdots+\mu_1+\nu)/a(\nu) = A(\mu_N+\cdots+\mu_1,\nu); \end{aligned} \tag{23}$$

the latter, when multiplied by the Liouville differintegration D^{ν}:

$$\prod_{k=1}^{N} A(\mu_k, \mu_{k-1}+\cdots+\mu_1+\nu)\frac{D^{\nu}F}{Dz^{\nu}} = A(\mu_N+\cdots+\mu_1,\nu)\frac{D^{\nu}F}{Dz^{\nu}}, \tag{24}$$

proves, by (20), the iteration rule for branchpoint operators.

PROPERTY 2. The iterated branchpoint operators of orders $\mu_1,\cdots,\mu_N$, applied to a Liouville differintegration of order ν, is equal to the Liouville differintegration of order $\mu_1+\cdots+\mu_N+\nu$:

$$\left(\frac{\delta}{\delta z}\right)^{\mu_N}\cdots\left(\frac{\delta}{\delta z}\right)^{\mu_1}\left(\frac{D^{\nu}F}{Dz^{\nu}}\right) = A(\mu_N+\cdots+\mu_1,\nu)\left(\frac{D}{Dz}\right)^{\mu_N+\cdots+\mu_1+\nu}F(z), \tag{25}$$

with a factor given by (21) with $\mu\equiv\mu_1+\cdots+\mu_N$: formula (24) assumes that ν is not an integer, and none of the $\mu_1+\nu,\cdots,\mu_N+\cdots+\mu_1+\nu$ are negative integers.

Remark 15. Although the coefficient of (25) is similar to that of (20) with $\mu\equiv\mu_1+\cdots+\mu_N$, the restrictions on the parameters are not. The condition $\mu+\nu$ not a negative integer is necessary, but not sufficient, in (25); there are further $(N-1)$

intermediate restrictions, namely, $\mu_1+\cdots+\mu_k+\nu$ not a negative integer, for $k=1,\cdots,N-1$.

Substituting $\mu\equiv\mu_1+\cdots+\mu_N$ in (20), viz.:

$$\left(\frac{\delta}{\delta z}\right)^{\mu_N+\cdots+\mu_1}\left(\frac{D^\nu F}{Dz^\nu}\right)=A(\mu_N+\cdots+\mu_1,\nu)\left(\frac{D}{Dz}\right)^{\mu_N+\cdots+\mu_1+\nu}F(z), \tag{26}$$

and comparing with (20), we follow the rule of composition of branchpoint operators.

PROPERTY 3. The iterated branchpoint operators of orders $\mu_1,\cdots,\mu_N$ are equivalent to a single branchpoint operator of order $\mu\equiv\mu_1+\cdots+\mu_N$, applied to a Liouville differintegration D^ν:

$$\left(\frac{\delta}{\delta z}\right)^{\mu_N}\cdots\left(\frac{\delta}{\delta z}\right)^{\mu_1}\left(\frac{D^\nu F}{Dz^\nu}\right)=\left(\frac{\delta}{\delta z}\right)^{\mu_N+\cdots+\mu_1}\left(\frac{D^\nu F}{Dz^\nu}\right), \tag{27}$$

assuming that ν is not an integer, and none of the $\mu_1+\nu,\cdots,\mu_N+\cdots+\mu_1+\nu$ are negative integers.

The branchpoint operator satisfies the same composition rule (27) as the Liouville differintegration, although the two operators are generally distinct. From (20) they coincide only if the coefficient is unity $A(\mu,\nu)=1$, which is the case in (21), with ν not an integer, if and only if μ is an integer. Thus we have the following degenerate rule.

PROPERTY 4. The branchpoint operator δ^μ coincides with the Liouville differintegration D^μ if and only if the order μ is an integer, i.e., both are ordinary derivatives or primitives:

$$\delta^\mu\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/\delta z^\mu=\frac{D^{\mu+\nu}F}{Dz^{\mu+\nu}}\leftrightarrow\mu\in Z, \tag{28}$$

where ν is not an integer.

Remark 16. The branchpoint and differintegration operators would have to coincide for ordinary derivations or integrations. The theorem shows that they are distinct generalizations, for all complex, nonintegral orders μ.

The degenerate rule can be extended to iterated branchpoint operators by using (25) instead of (20), viz., Property 5.

PROPERTY 5. The iterated branchpoint operators $\delta^{\mu_N}\cdots\delta^{\mu_1}$ of orders $\mu_N,\cdots,\mu_1$ coincide with the Liouville differintegration D^μ of order equal to the sum $\mu\equiv\mu_1+\cdots+\mu_N$ if and only if the latter is an integer:

$$\left(\frac{\delta}{\delta z}\right)^{\mu_N}\cdots\left(\frac{\delta}{\delta z}\right)^{\mu_1}\left(\frac{D^\nu F}{Dz^\nu}\right)=\left(\frac{D}{Dz}\right)^{\mu_N+\cdots+\mu_1+\nu}F(z)\leftrightarrow\mu\equiv\mu_1+\cdots+\mu_N\in\mathbb{Z}, \tag{29}$$

where ν is not an integer, and none of $\mu_1+\nu,\cdots,\mu_N+\cdots+\mu_1+\nu$ is a negative integer.

Remark 17. Property A5 follows by applying A3 to A4, i.e., (27) and (28) yield (29).

Remark 18. The result (29) shows that the composition of branchpoint operators $\delta^{\mu_1},\cdots,\delta^{\mu_N}$, with complex nonintegral orders $\mu_1,\cdots,\mu_N$, which are not Liouville differintegrations by property A4, may be equivalent to a Liouville differintegration of order $\mu\equiv\mu_1+\cdots+\mu_N$, if and only if the latter is an integer.

Property 4, in the form of Remark 16, that the branchpoint and differintegration operators do not coincide for nonintegral complex orders, opens the way for the annihilation of the latter by the former (since they must be applied in this order, this is the only annihilation possible). To find the cases, i.e., properties A4 and A5, of coincidence of branchpoint and differintegration operators, we have required that the coefficient (21) be unity; the annihilation of the latter by the former generally requires

the coefficient (21) to vanish. This is possible, for nonintegral complex ν, if and only if $\mu+\nu$ is an integer, i.e., $\mu=p-\nu$ where p is any integer $p\in\mathbb{Z}$. We have proved the following annihilation property.

PROPERTY 6. The Liouville differintegration D^ν of nonintegral complex order ν, is annihilated by the branchpoint operator δ^μ, if and only if its order $\mu=p-\nu$ differs from $-\nu$ by an integer p:

$$\nu\in\mathbb{C}-\mathbb{Z},\ p\in\mathbb{Z}\colon\ \delta^{p-\nu}\left(\frac{D^\nu F}{Dz^\nu}\right)\Big/\delta z^{p-\nu}=0. \tag{30}$$

Remark 19. All Liouville differintegrations D^ν can be annihilated by a suitable branchpoint operator, except the function itself $\nu=0$, an ordinary derivative $\nu\in\mathbb{N}$, or an ordinary primitive $-\nu\in\mathbb{N}$.

Remark 20. For each Liouville differintegration D^ν of complex nonintegral order ν, there is a denumerable infinity of branchpoint operators $\delta^{p-\nu}$ with integer p that annihilates it.

Property A6 extends to annihilation by iterated branchpoint operators.

PROPERTY 7. The iterated branchpoint operators $\delta^{\mu_N},\cdots,\delta^{\mu_1}$ of orders $\mu_N,\cdots,\mu_1$ annihilate a Liouville differintegration D^ν of order ν, if and only if the sum of all orders is an integer:

$$\left(\frac{\delta}{\delta z}\right)^{\mu_N}\cdots\left(\frac{\delta}{\delta z}\right)^{\mu_1}\left(\frac{D^\nu F}{Dz^\nu}\right)=0\leftrightarrow\mu_N+\cdots+\mu_1+\nu\in\mathbb{Z}, \tag{31}$$

where ν is not an integer, and none of $\mu_1+\nu,\cdots,\mu_N+\cdots+\mu_1+\nu$ are negative integers.

Remark 21. The use of iterated branchpoint operators (31) increases the number of possibilities of annihilating any given Liouville differintegration of nonintegral complex order ν, relative to that stated in Remark 19; i.e., we have a continuum of possibilities instead of a denumerable set.

Remark 22. If we exclude the possibility of annihilating functions and ordinary derivatives and primitives, then the branchpoint operator goes as far as we could expect, i.e., (i) it can annihilate every other instance of the Liouville differintegration, and (ii) it offers an infinity of possibilities of doing so.

5. Discussion. Although the Liouville differintegration and branchpoint operator are generally distinct, they are both derivation operators since they are linear and satisfy the Leibnitz rule, in the extended form [17], [18], [2], [4], which we quote first.

THEOREM 6. *The Liouville differintegration of order ν, of the product of two analytic functions $F(z)$, $G(z)$, satisfies the generalized Leibnitz rule*:

$$\frac{D^\nu\{F(z)G(z)\}}{Dz^\nu}=\sum_{k=0}^{\infty}\binom{\nu}{k}G^{(k)}z\frac{D^{\nu-k}F}{Dz^{\nu-k}} \tag{32}$$

where $F(z)$, and $F(z)G(z)$ satisfy the asymptotic condition (3).

Proof. Since the proof differs from those in the literature in matters of detail, we give a brief outline here. The asymptotic condition (3) for $F(z)G(z)$ implies that we need consider only part L of the Hankel path in (1) that lies within a circle of finite radius $|\zeta|\leqq R$; the remainder of the path gives a term $O(R^{-\varepsilon})\to 0$ uniformly as $R\to\infty$. Thus we have

$$\frac{D^\nu\{F(z)G(z)\}}{Dz^\nu}=\{\Gamma(1+\nu)/2\pi i\}\int_L F(\zeta)G(\zeta)(\zeta-z)^{-\nu-1}\,d\zeta. \tag{33}$$

The function $G(z)$ is analytic in the circle $|\zeta| \leqq R+\delta$ for some $\delta > 0$, and thus has the Taylor series

$$G(\zeta) = \sum_{k=0}^{\infty} \{(\zeta - z)^k / k!\} G^{(k)}(z), \tag{34}$$

uniformly convergent in a closed subcircle $|\zeta| \leqq R$. The latter (34) may be substituted into (33), and integrated term by term:

$$\frac{D^\nu\{F(z)G(z)\}}{Dz^\nu} = \sum_{k=0}^{\infty} \{\Gamma(1+\nu)/k!2\pi i\} G^{(k)}(z) \int_L F(\zeta)(\zeta - z)^{-\nu+k-1}\, d\zeta. \tag{35}$$

Since $F(z)$ satisfies the asymptotic condition (3) with ε, it also satisfies $\varepsilon - k$ for all positive integer k, and path L may be extended back to the Hankel loop:

$$\begin{aligned} \frac{D^\nu\{F(z)G(z)\}}{Dz^\nu} &= \sum_{k=0}^{\infty} \{\Gamma(1+\nu)/k!\Gamma(1+\nu-k)\} G^{(k)}(z) \\ &\quad \cdot \{\Gamma(1+\nu-k)/2\pi i\} \int_{\infty \exp\{i \arg(z)\}}^{(z+)} F(\zeta)(\zeta - z)^{-\nu+k-1}\, d\zeta; \end{aligned} \tag{36}$$

applying (1), we obtain

$$\frac{D^\nu\{F(z)G(z)\}}{Dz^\nu} = \sum_{k=0}^{\infty} \{\Gamma(1+\nu)/k!\Gamma(1+\nu-k)\} G^{(k)}(z) \frac{D^{\nu-k}F}{Dz^{\nu-k}}, \tag{37}$$

which is equivalent to (32), with the notation

$$\binom{\nu}{k} \equiv \Gamma(1+\nu)/\Gamma(1+k)\Gamma(1+\nu-k). \tag{38}$$

We have proved the extended Leibnitz rule. □

We may now proceed to prove that δ^μ is a differentiation operator, by the following sequence of reasoning.

Remark 23. In the case where $\nu = n$ is a positive integer, $(\nu - k)! = \infty$ for $k > n$, and (32) reduces to the ordinary Leibnitz rule:

$$\frac{D^n\{F(z)G(z)\}}{Dz^n} = \sum_{k=0}^{n} \binom{m}{k} F^{(n-k)}(z) G^{(k)}(z). \tag{39}$$

Remark 24. The ordinary Leibnitz rule (33) looks like a binomial expansion of derivates, and the extended Leibnitz rule (32) looks like a binomial series of Liouville differintegrations.

THEOREM 7. *The Liouville differintegration D^ν is a derivation operator, because* (i) *it is linear, and* (ii) *it satisfies the extended Leibnitz rule.*

Proof. (i) The linear property results from Definition 1 as a generalized Cauchy integral (1).

(ii) The extended Leibnitz rule is proved in Theorem 6.

A similar result holds for the branchpoint operator.

PROPERTY 8. The branchpoint operator δ^μ is a derivation operator, since (i) it is linear, and (ii) it satisfies the extended Leibnitz rule in the following form:

$$\delta^\mu \left\{ \frac{D^\nu(F(z)G(z))}{Dz^\nu} \right\} \Big/ \delta z^\mu = \sum_{k=0}^{\infty} \binom{\mu+\nu}{k} G^{(k)}(z) \delta^{\mu-k} \left(\frac{D^\nu F}{Dz^\nu} \right) \Big/ \delta z^{\mu-k}, \tag{40}$$

where μ, ν are not integers.

Proof. The linear property (i) of the branchpoint operator δ^{μ} follows from that of the Liouville differintegration because the factor (21) connecting them (20) does not involve the independent z or dependent $F(z)$ $G(z)$ variables.

The factor (21) is unchanged if we add to the arguments any integer numbers p, q, viz.,

$$p, q \in \mathbb{Z}: A(\mu+p, \nu+q) = A(\mu, \nu). \tag{41}$$

Writing the Leibnitz rule (32) with ν replaced by $\mu+\nu$, and multiplying by (41), we obtain

$$A(\mu, \nu)\frac{D^{\mu+\nu}\{F(z)G(z)\}}{Dz^{\mu+\nu}} = \sum_{k=0}^{\infty}\binom{\nu}{k}G^{(k)}(z)A(\mu-k, \nu)\frac{D^{\mu+\nu-k}F}{Dz^{\mu+\nu-k}}; \tag{42}$$

substituting (20) into (42), yields (40), which is (ii), the extended Leibnitz rule for the branchpoint operator. We must exclude integer ν, and also the negative integer μ, $\mu-1, \cdots, \mu-k, \cdots$; the latter set of conditions is equivalent to stating that μ is not an integer. □

As an example of the application of the branchpoint δ^{μ} and differintegration D^{μ} operators to special functions, we consider the Hermite function.

DEFINITION 6. The Hermite function $H_{\nu}(z)$ of complex order ν and variable z is defined by Liouville differintegration with order ν of the Gaussian function of variable z:

$$H_{\nu}(z) \equiv e^{i\pi\nu}e^{z^2}\frac{D^{\nu}(e^{-z^2})}{Dz^{\nu}}. \tag{43}$$

Remark 25. In the case where $\nu = n$ is a positive integer, Definition (43) agrees with that of the Hermite [7] polynomial:

$$H_n(z) = (-)^n e^{z^2}\{e^{-z^2}\}^{(n)}. \tag{44}$$

Remark 26. Using (1), the Hermite function of complex ν, nonnegative integral order, has the following integral representation:

$$H_{\nu}(z) = \{\Gamma(1+\nu)/2\pi i\}\, e^{z^2}\int_{\infty \exp\{i \arg(z)\}}^{(z+)} (z-\zeta)^{-\nu-1}\, e^{-\zeta^2}\, d\zeta, \tag{45}$$

along a Hankel loop (Fig. 1).

Remark 27. When we use (2), the Hermite function of complex order ν, with negative real part $\mathrm{Re}(\nu) < 0$, has the integral representation

$$H_{\nu}(z) = \{e^{z^2}/\Gamma(-\nu)\}\int_{\infty \exp\{i \arg(z)\}}^{z} (x-z)^{-\nu-1}\, e^{x^2}\, dx, \tag{46}$$

along a semi-infinite straight line (Fig. 1).

Remarks 24–26 show that the definition of the Hermite function via the Liouville differintegration (43) agrees with other definitions in the literature [5], and applies to all complex orders ν.

We proceed to prove the annihilation property.

PROPERTY 9. The Hermite function (43) of nonintegral complex order ν, multiplied by a Gaussian function e^{-z^2} of the independent variable z, is annihilated by the branchpoint operator $\delta^{p-\nu}$ of order $p-\nu$, differing from $-\nu$ by an integer p:

$$\nu \in \mathbb{C}-\mathbb{Z}, \quad p \in \mathbb{Z}: \delta^{p-\nu}\{e^{-z^2}H_{\nu}(z)\}/\delta z^{p-\nu} = 0. \tag{47}$$

Remark 28. The need for the Gaussian function e^{-z^2} as a factor in the curly brackets in (47), follows from (30), the general annihilation Property 5, in the particular form:

$$0 = e^{i\pi\nu}\delta^{p-\nu}\left\{\frac{D^{\nu}(e^{-z^2})}{Dz^{\nu}}\right\}\Big/\delta z^{p-\nu} = \delta^{p-\nu}\{e^{-z^2}H_{\nu}(z)\}/\delta z^{p-\nu}. \tag{48}$$

Proof. Remark 28 may serve as a short, indirect proof of (47). We may also prove Property 9 directly, by substituting (48) into Definition 4 of branchpoint operator (6):

$$\begin{aligned}\delta^{p-\nu}\left\{\frac{D^{\nu}(e^{-z^2})}{Dz^{\nu}}\right\}\Big/\delta^{p-\nu} &\equiv \{\Gamma(1+\nu)/2\pi i\}\int_{\infty\exp\{i\arg(z)\}}^{(z+)}\frac{d^{p-\nu}\{(\zeta-z)^{-\nu-1}\}}{dz^{p-\nu}}\,e^{-\zeta^2}\,d\zeta,\\ &= \{\Gamma(-\nu)\Gamma(1+\nu)/2\pi i\Gamma(-p)\}\,e^{i\pi(p-\nu)}\cdot\int_{\infty\exp\{i\arg(z)\}}^{(z+)}(\zeta-z)^{-p-1}\,e^{-\zeta^2}\,d\zeta \qquad (49)\\ &= \{\Gamma(-\nu)\Gamma(1+\nu)/\Gamma(-p)\Gamma(1+p)\,e^{i\pi\nu}\}H_p(z),\end{aligned}$$

where (45) was used. The Hermite polynomial in (49) does not vanish, but the coefficient in curly brackets does:

$$C(\nu,p)\equiv\Gamma(-\nu)\Gamma(1+\nu)/\Gamma(-p)\Gamma(1+p) = \sin(\pi p)/\sin(\pi\nu) = 0, \tag{50}$$

because p is an integer. Substitution of (50) into (49) proves (48) and (47). □

If in (47) we set $\mu \equiv p-\nu$, and replace the branchpoint operator δ^{μ} with a Liouville differintegration, the result is not zero, but rather, as follows from (43) and (11):

$$\begin{aligned}\frac{D^{\mu}\{e^{-z^2}H_{\nu}(z)\}}{Dz^{\mu}} &= e^{i\pi\nu}D^{\mu}\left\{\frac{D^{\nu}(e^{-z^2})}{Dz^{\nu}}\right\}\Big/Dz^{\mu}\\ &= e^{i\pi\nu}D^{\mu+\nu}(e^{-z^2})Dz^{\mu+\nu} = e^{-i\pi\mu}\,e^{-z^2}H_{\mu+\nu}(z),\end{aligned} \tag{51}$$

a Hermite function of order $\mu+\nu$. The result (51) can be stated as an integral identity.

THEOREM 8. *The Hermite function satisfies the integral identities*:

$$e^{-z^2}H_{\mu+\nu}(z) = \{\Gamma(1+\mu)/2\pi i\}\int_{\infty\exp\{i\arg(z)\}}^{(z+)}(z-\zeta)^{-\mu-1}H_{\nu}(\zeta)\,e^{-\zeta^2}\,d\zeta \tag{52}$$

$$= \{\Gamma(-\mu)\}^{-1}\int_{\infty\exp\{i\arg(z)\}}^{z}(\zeta-x)^{-\mu-1}H_{\nu}(x)\,e^{-x^2}\,dx, \tag{53}$$

where the Hankel path (52) *applies for* μ *not a negative integer, and the straight path* (53) *for* $\operatorname{Re}(\mu)<0$.

Proof. Substitution of (1) and (2) into (51), yields, respectively, (52) and (53).

The integral identities (52) and (53) allow the evaluation of certain integrals of Hermite functions in terms of other Hermite functions. They are similar to results holding for other special functions [2].

REFERENCES

[1] L. M. B. C. CAMPOS, *On a concept of derivative of complex order with application to special functions*, IMA J. Appl. Math., 33 (1984), pp. 109-133.

[2] ———, *On rules of derivation with complex order of analytic and branched functions*, Portugal. Math., 43 (1985), pp. 347-376.

[3] ———, *On a systematic approach to some properties of special functions*, IMA J. Appl. Math., 36 (1986), pp. 191-206.

[4a] L. M. B. C. CAMPOS, *On extensions of Laurent's theorem in the fractional calculus, with application to the generation of higher transcendental functions*, Mat. Vesnik, 38 (1986), pp. 375–390.

[4b] ———, *Erratum*, Mat. Vesnik, 40 (1988).

[5] A. ERDELYI, ed., *Higher Transcendental Functions*, 3 vols., McGraw-Hill, New York, 1953.

[6] H. HANKEL, *Die Euler'schen Integrale bei unbeschrankter Variabilitat des Argumentes*, Z. Math. Phys., 9 (1864), pp. 7–11.

[7] C. HERMITE, *Sur un nouveau developpement en série de fonctions*, Compt. Rend. Acad. Sci. Paris, 58 (1884) pp. 83–100 and 266–273; Oeuvres 2 (1884), pp. 293–302.

[8] J. L. LAVOIE, T. J. OSLER, AND R. TREMBLAY, *Fractional derivatives and special functions*, SIAM Rev., 18 (1976), pp. 240–268.

[9] J. L. LAVOIE, R. TREMBLAY, AND T. J. OSLER, *Fundamental properties of fractional derivatives via Pochhammer integrals*, Ross (1974), pp. 323–356.

[10] A. V. LETNIKOV, *Theory of differentiation of fractional order*, Mat. Sb., 3 (1868), pp. 1–68.

[11] J. LIOUVILLE, *Mémoire sur le calcul des differentielles à indices quelconques*, J. Éc. Polyt., 13 (1832), pp. 71–162.

[12] A. C. MCBRIDE, *Fractional Calculus and Integral Transforms of Generalized Functions*, Pitman, Boston, 1979.

[13] A. C. MCBRIDE AND G. F. ROACH, *Fractional Calculus*, Pitman, Boston, 1986.

[14] P. A. NEKRASSOV, *Generalized differentiation*, Mat. Sb., 14 (1888), pp. 45–168.

[15] K. NISHIMOTO, *Fractional Calculus*, 2 vols., Descartes Press, Koryama, Japan, 1984.

[16] K. B. OLDHAM AND J. SPANIER, *Fractional Calculus*, Academic Press, New York, 1974.

[17] T. J. OSLER, *Leibnitz rule for fractional derivatives generalized, and an application to infinite series*, SIAM J. Appl. Math., 18 (1970), pp. 658–674.

[18] ———, *A further extension of the Leibnitz rule to fractional derivatives and its relation to Parseval's formula*, SIAM J. Math. Anal., 3 (1972), pp. 1–16.

[19] C. POCHHAMMER, *Uber ein Integral mit doppeltem Unlauf.*, Math. Ann., 75 (1890), pp. 470–494.

[20] B. RIEMANN, *Versuch einer Allgemeinen Aufassung der Integration und Differentiation*, Ges. Werke, (1847), pp. 331–344.

[21] B. ROSS, ed., *Fractional Calculus and Applications*, Springer-Verlag, Berlin, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 454-467, March 1989

013

ON A GENERALIZED MITTAG-LEFFLER THEOREM AND IMPLICIT DIFFERINTEGRATION*

L. M. B. C. CAMPOS†

Abstract. The differintegration operator is defined in the literature [B. Ross, ed., *Fractional Calculus and Applications*, Springer-Verlag, Berlin, New York, 1974; *Fractional Calculus*, Academic Press, New York, 1974; *Fractional Calculus*, 2 vols., Descartes Press, New York, 1974; *IMA J. Appl. Math.*, 33 (1984), pp. 109-133] as an extension to real or complex order ν, of the ordinary nth derivative and primitive, which correspond to, respectively, $\nu = +n$, $-n$, a positive, negative integer. The differintegration operator has been used to generalize [*SIAM J. Math. Anal.*, 1 (1970), pp. 288-293; *SIAM J. Math. Anal.*, 2 (1971), pp. 37-48; *Mat. Vesnik*, 38 (1986), pp. 375-390; Mat. Vesnik, 40 (1988), p. 85] the power series associated with the names of Taylor, Laurent, Lagrange, and Teixeira. Here we generalize the Mittag-Leffler theorem on series of fractions, as a representation of the differintegration of a meromorphic function (§ 3); the proof uses the principal parts of the Laurent series expansion, and of the differintegration of a function near a pole (§ 2). The rules of derivation with complex order have been discussed in the literature [*SIAM J. Appl. Math.*, 18 (1970), pp. 658-674; *SIAM J. Math. Anal.*, 3 (1972), pp. 1-16; *Portugal Math.*, 43 (1985), pp. 347-376], and we prove (§ 3) a new rule of differintegration of implicit functions. The differintegration operator has many applications to special functions [Ross, op. cit., pp. 323-356; *SIAM Rev.*, 18 (1976), pp. 240-268; *IMA J. Appl. Math.*, 36 (1986), pp. 191-206], and we illustrate our results in this regard (§ 5) with Hermite functions.

Key words. implicit differintegration, series of fractions, meromorphic functions, singularities, residues, Hermite functions

AMS(MOS) subject classifications. 30A99, 30B99, 30D30, 30E20, 33A65

1. Introduction.

The differintegration operator has been defined in the context of real [23], [18], generalized [14], [15], and complex [17], [1] functions, as an extension of the ordinary derivatives and primitives. It has some properties analogous to those of ordinary derivatives, e.g., if $F(z)$ is an analytic function, its Liouville differintegration $D^\nu F/Dz^\nu$ with $\mathrm{Re}\,(\nu) > 0$ is also analytic (§ 2). This extends the known result that an analytic function is infinitely differentiable from $\nu = n$ positive integer to the whole right-hand ν-half-plane. It suggests that the differintegration D^ν of an analytic function $F(z)$ has a Taylor series type of expansion in ascending powers of z. If the function $F(z)$ has an isolated singularity at $z = \zeta$, its differintegration $D^\nu F/Dz^\nu$ has a Laurent series type of expansion, involving ascending and descending powers of $z - \zeta$. It is not the purpose of this paper to address the representation of the differintegration of functions into power series, which is discussed elsewhere in the literature [19], [21], [3]. For our purposes, it is sufficient to note that the Laurent series type of expansion specifies the principal part of the differintegration of a function $D^\nu F/Dz^\nu$ near singularity, either essential or pole.

The latter remark suggests a form of representation of the differintegration of a rational function; such a function is analytic except for a finite number of poles, and the principal parts near these specify its differintegration to within an added constant, which may be evaluated at any regular point. This process can be extended to a function with an infinite number of poles and no essential singularities, i.e., such that the poles "spread" to infinity, and do not "cluster" in the neighborhood of an accumulation point (which would be an essential singularity). This leads to a generalization of the Mittag-Leffler theorem [24], by representing the differintegration of a meromorphic

* Received by the editors December 1, 1987; accepted for publication (in revised form) May 6, 1988. This was supported by a J.N.I.C.T. research project and by CAUTL, Instituto de Fisica-Matemática, I.N.I.C.

† Instituto Superior Técnico, 1096 Lisboa Codex, Portugal.

function as a series of fractions, viz., the principal parts at the poles. The added constant can be evaluated at any regular point, provided that the series converges uniformly in its neighborhood. The question of convergence is the only point where the proof of the generalized Mittag-Leffler theorem for the differintegration of a meromorphic function goes beyond the analogous finite expansion for a rational function.

Since the differintegration is a generalization of the ordinary derivative, the question arises of extending the rules of derivation to complex order, e.g., differintegration of elementary functions [10], [2]. The Leibnitz rule can also be extended [20], [22], [2] into an infinite chain rule, which shows that several important special functions are differintegrations of products of elementary functions [11], [2]. Here we prove (§ 4) a rule of implicit differintegration of a function with regard to another; this rule may be combined with the chain rule, and both lead to series expansions. The use of Riemann differintegration to obtain properties of special functions, both short proofs of known results and new formulas, is well documented in the literature [18], [6]. The Riemann [10] rather than the Liouville [17] differintegration should be used for hypergeometric, confluent, Legendre, Laguerre, Jacobi, Gegenbauer, and Chebyshev functions, because these involve branchpoints. The Hermite function involves the Liouville differintegration of an integral function, namely [5] the Gaussian, and we choose it (§ 5) for examples of application of rules of differintegration, since it appears to have received less attention in the literature.

2. Extensions of the Laurent series and calculation of residues. We recall the definition of the Liouville [13] system of differintegration, in the context [12], [16], [17], [1] of functions of a complex variable.

DEFINITION 1. The differintegration D^ν of complex order ν, not a negative integer, of an analytic function $F(z)$ is defined by the generalized Cauchy integral:

$$\frac{D^\nu F}{Dz^\nu} = \{\Gamma(1+\nu)/2\pi i\} \int_{\infty \exp\{i \arg(z)\}}^{(z+)} (\zeta - z)^{-\nu-1} F(\zeta)\, d\zeta, \tag{1}$$

along a Hankel [8] loop (Fig. 1) going round $\zeta = z$ in the counterclockwise direction, and starting and ending at infinity, respectively, above and below the branchcut $C \equiv \{\zeta: |\zeta| > |z|, \arg(\zeta) = \arg(z)\}$. The function satisfies the following asymptotic condition:

$$\arg(\zeta) - \delta < \arg(z) < \arg(\zeta) + \delta: F(\zeta) \sim O(\zeta^{\mathrm{Re}(\nu)-\varepsilon}), \tag{2}$$

for some $\varepsilon, \delta > 0$ in a sector about the branchcut C, so that the improper integral in $d\zeta$ converges (1). It is assumed that the convergence is uniform with regard to z in a region D.

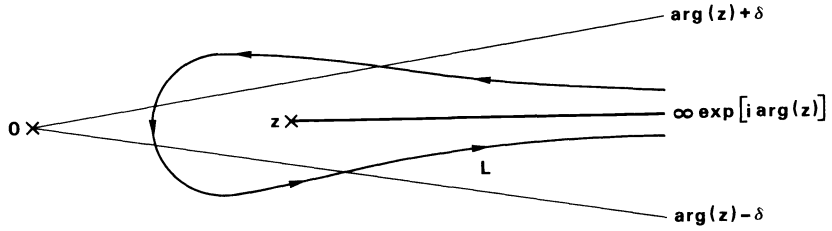


FIG. 1. *The differintegration of an analytic function involves a contour integral* (1) *along the Hankel* [8] *path, going counterclockwise around the branchpoint* $\zeta = z$, *and starting and ending at infinity in the* ζ*-plane, respectively, above and below the semi-infinite branchcut, joining* z *to infinity in the direction of* arg (z), *so that in an angular sector containing the branchcut the function satisfies an asymptotic condition* (2).

Remark 1. The extension of Definition 1 to negative integer order ν, the relation with ordinary derivatives and primitives, and the integration with complex order are summarized [5] and proved [1] elsewhere. For the present we note as an immediate consequence of Definition 1 that the asymptotic condition (2), for the convergence of (1), is still met with ν replaced by μ, such that $\mathrm{Re}(\mu) \geqq \mathrm{Re}(\nu)$, thus proving Theorem 1.

THEOREM 1. *If the differintegration $D^\nu F/Dz^\nu$ of an analytic function $F(z)$, with complex order ν exists, then the differintegration $D^\mu F/Dz^\mu$ with complex order μ, such that* $\mathrm{Re}(\mu) \geqq \mathrm{Re}(\nu)$, *also exists.*

Remark 2. If we set $\mu \neq \nu$ and $\mathrm{Re}(\mu) = \mathrm{Re}(\nu)$, i.e., $\mathrm{Im}(\mu) \neq \mathrm{Im}(\nu)$, Theorem 1 still holds, showing that the imaginary part of the order of differintegration $D^\nu F$ of a function $F(z)$ does not affect existence.

If we set $\mu = \nu+1, \cdots, \nu+n, \cdots$ in Theorem 1, it follows that $D^\nu F/Dz^\nu$, if it exists, is infinitely differentiable, and thus, in the context of complex analysis, specifies an analytic function.

THEOREM 2. *The differintegration D^ν (1) of (2) an analytic function $F(z)$, if it exists, specifies a function $G(z) \equiv D^\nu F/Dz^\nu$, which is analytic in the region D, of uniform convergence of the integral* (1).

An analytic function has a Taylor series expansion, which we may expect to hold for the function $G(z) \equiv D^\nu F/Dz^\nu$ in Theorem 2. Noting that $G^{(k)}(z) \equiv D^{\nu+k}F/Dz^{\nu+k}$ by the rule of composition of derivatives [1], [5], we may expect the following theorem to hold.

THEOREM 3 (extended Taylor series). *The differintegration D^ν, with complex order ν, of an analytic function $F(z)$, can be represented by an extended Taylor series of ascending powers*:

$$\frac{D^\nu F}{Dz^\nu} = \sum_{k=0}^{\infty} \{(z-\zeta)^k/k!\} \frac{D^{\nu+k}F}{D\zeta^{\nu+k}}, \tag{3}$$

whose coefficients involve differintegrations of orders $\nu, \nu+1, \cdots, \nu+k, \cdots$, equal to ν plus a nonnegative integer, at a regular point ζ. If η is the singularity of $F(z)$ closest to the regular point ζ, and $R \equiv |\eta-\zeta|$, then the series (3) *converges:* (i) *absolutely, in the open circle $|z-\zeta| < R$*; (ii) *uniformly, in the closed subcircle $|z-\zeta| \leqq R-\varepsilon$, for some ε, such that $0 < \varepsilon < R$.*

Remark 3. The regions of absolute and uniform convergence of the extended Taylor series (3) are the same as for the original Taylor series, which is obtained by setting $\nu = 0$, in which case only ordinary derivatives are needed.

Proof. The proof of (3) is broadly similar to the standard demonstration of the Taylor series [24], except that the binomial rather than the geometric series is needed to establish the extended Taylor series [4]. The details are too extensive [6] to be reproduced here.

If the function $F(z)$ has an isolated singularity at $z = \zeta$, it can be expanded in the neighborhood in a Laurent series [24], and its differintegration D^ν with complex order has an extended Laurent series [6], viz., Theorem 4.

THEOREM 4 (extended Laurent series). *The differintegration D^ν, with complex order ν, of a function $F(z)$, having an isolated singularity at $z = \zeta$, can be represented by an extended Laurent series*:

$$\frac{D^\nu F}{Dz^\nu} = \sum_{k=0}^{\infty} A_k (z-\zeta)^k + \sum_{k=1}^{\infty} A_{-k} (z-\zeta)^{-\nu-k}, \tag{4}$$

where the coefficients A_k of the ascending integral powers, and A_{-k} of the descending

complex powers, are given, respectively, by

$$A_k \equiv \{\Gamma(1+\nu+k)/k!2\pi i\} \int_L (z-\zeta)^{-\nu-k-1} F(z)\, dz, \tag{5}$$

$$A_{-k} \equiv e^{i\pi\nu}\{\Gamma(\nu+k)/(k-1)!2\pi i\} \int_l (z-\zeta)^{k-1} F(z)\, dz, \tag{6}$$

where L, l are Hankel paths like (1), *with l lying within L. If* $z=\eta$ *is the singularity of* $F(z)$ *closest to* $z=\zeta$, *then the series* (4) *converges*: (i) *absolutely in the open disk* $0<|z-\zeta|<R$ *of radius* $R\equiv|\zeta-\eta|$; (ii) *uniformly in the closed subdisk* $\varepsilon \leqq |z-\zeta| \leqq R-\delta$, *with* $\varepsilon, \delta > 0$ *and* $\varepsilon+\delta<R$.

Remark 4. In the absence of differintegration $\nu=0$, the factors in curly brackets in (5), (6) reduce to $(2\pi i)^{-1}$, and only integral powers appear in (4), which reduces to the original Laurent series.

Remark 5. If the function $F(z)$ is analytic at $z=\zeta$, then the inner path l can be shrunk to zero, so that (6) vanishes and (4) reduces to the first sum, where the coefficients (5) can be evaluated using (1):

$$A_k = (k!)^{-1} D^{\nu+k}F/D\zeta^{\nu+k}, \tag{7a}$$

$$A_{-k} = 0. \tag{7b}$$

Substituting (7a), (7b) into (4), we regain the extended Taylor series (3).

Remark 6. The regions of absolute and uniform convergence of the extended Laurent series (5), (6) are the same as for the original Laurent series with $\nu=0$, because they are independent of ν; they differ from the regions of absolute and uniform convergence of the Taylor series (extended (3) and original $\nu=0$) only in excluding, respectively, an open (closed) neighborhood of the singularity $z=\zeta$.

The descending powers of the extended Laurent series, viz., the second term on the right-hand side of (4), specify the principal part of the differintegration $D^\nu F/Dz^\nu$ in the neighborhood of the singularity, viz., Property 1.

PROPERTY 1 (principal part of a differintegration). The differintegration D^ν, with complex order ν, of a function $F(z)$, with an isolated singularity at $z=\zeta$, has principal part consisting of: (i) an unending sequence of powers with exponents $-\nu-1, \cdots, -\nu-k, \cdots$:

$$\frac{D^\nu F}{Dz^\nu} \sim A_{-1}(z-\zeta)^{-\nu-1} + A_{-2}(z-\zeta)^{-\nu-2} + \cdots, \tag{8}$$

if $z=\zeta$ is an essential singularity; (ii) a sequence of m powers of exponents $-\nu, -\nu-1, \cdots, -\nu-m$;

$$\frac{D^\nu F}{Dz^\nu} \sim A_{-1}(z-\zeta)^{-\nu-1} + \cdots + A_{-m}(z-\zeta)^{-\nu-m}, \tag{9}$$

if $z=\zeta$ is a pole of order m.

The coefficient A_{-1} is the residue and can be calculated, at a pole, from the differintegration (9), as follows.

PROPERTY 2. The residue A_{-1} of the differintegration $D^\nu F/Dz^\nu$ at a pole of order m (9), can be calculated by

$$A_{-1} = \lim_{z\to\zeta} \{(m-1)!\}^{-1} \left(\frac{d}{dz}\right)^{m-1} \{(z-\zeta)^{m+\nu} D^\nu F/Dz^\nu\} \tag{10}$$

where $(d/dz)^{m-1}$ is an ordinary derivative.

Remark 7. In the case $m=1$ of a simple pole, the residue (10) can be calculated without derivations, by

$$A_{-1}=\lim_{z\to\zeta}(z-\zeta)^{1+\nu}\frac{D^{\nu}F}{Dz^{\nu}}. \tag{11}$$

Not only the residue A_{-1}, but also all other coefficients A_{-k}, of principal part of the differintegration at a multiple pole (9), can be calculated by a formula analogous to (10).

PROPERTY 3. The coefficients A_{-k} with $k=1,\cdots,m$, of the principal part (9) of the differintegration at a pole of order m, can be calculated by

$$A_{-k}=\lim_{z\to\zeta}\{(m-k)!\}^{-1}\left(\frac{d}{dz}\right)^{m-k}\left\{(z-\zeta)^{m+\nu}\frac{D^{\nu}F}{Dz^{\nu}}\right\}. \tag{12}$$

Remark 8. The residue A_{-1} (10) is the particular case $k=1$ of (12).

3. The generalized Mittag–Leffler theorem for meromorphic functions. The identification of the principal part of a differintegration $D^{\nu}F/Dz^{\nu}$ at a singularity opens the way to the representation of functions having no singularities other than poles by sequences of fractions. For rational and meromorphic functions, which have, respectively, a finite and an infinite number of poles, the representation by means of regular or singular power series, e.g., Taylor (3) or Laurent (4)-(6) type, requires extensive use of analytic continuation because the radius of convergence of each series is limited by the nearest singularity. Also, the form of the coefficients tends to be complicated in such cases, e.g., they involve Bernoulli or Euler numbers for the compound trigonometric and hyperbolic functions, viz., the tangent, cotangent, secant, and cosecant, which are meromorphic. The representation by means of series of fractions is much more convenient, since a single expansion holds over the whole complex plane, and the coefficients are determined by the principal parts at the poles. We start with the case of rational functions, having a finite number M of poles, for which questions of convergence do not arise. We assume that the differintegration $D^{\nu}F/Dz^{\nu}$ is a rational function, with a finite number M of poles, at the points $z=a_k$, with orders b_k, for $k=1,\cdots,M$, with A^{j}_{-k} as coefficients of the principal parts:

$$\frac{D^{\nu}F}{Dz^{\nu}}\sim\sum_{j=1}^{b_k}A^{k}_{-j}(z-a_k)^{-\nu-j}; \tag{13}$$

the point at infinity may also be a pole, of order b_0, with coefficients A^{0}_{-j} of the principal part of the differintegration:

$$\frac{D^{\nu}F}{Dz^{\nu}}\sim\sum_{j=0}^{b_0}A^{0}_{-j}z^{\nu+j}. \tag{14}$$

Thus the function $G(z)$, obtained by subtracting from $D^{\nu}F/Dz^{\nu}$ its principal parts (14), (13) at all the poles, has no singularities in the extended complex plane and reduces to a constant:

$$G(z)\equiv\frac{D^{\nu}F}{Dz^{\nu}}-\sum_{j=0}^{b_0}A^{0}_{-j}z^{\nu+j}-\sum_{k=1}^{M}\sum_{j=1}^{b_k}A^{k}_{-j}(z-a_k)^{-\nu-j}=\text{const.}=G(\zeta), \tag{15}$$

e.g., its value at any regular point $z=\zeta$.

The result (15) may be stated formally as the following theorem.

THEOREM 5 (expansion in fractions). *If the differintegration $D^{\nu}F/Dz^{\nu}$ with complex order ν is a rational function, with M poles a_k of orders b_k, with $k=1,\cdots,M$ in the z-plane, with coefficients A^{k}_{-j} with $j=1,\cdots,b_k$ of the principal parts* (13), *plus possibly*

a pole of order b_0 at infinity, with coefficients A^0_{-j} with $j = 1, \cdots, b_0$, of the principal part (14), *then it can be represented by a finite expansion in fractions*:

$$\frac{D^\nu F}{Dz^\nu} = \frac{D^\nu F}{D\zeta^\nu} + \sum_{j=1}^{b_0} A^0_{-j}(z^{\nu+j} - \zeta^{\nu+j}) + \sum_{k=1}^{M} \sum_{j=1}^{b_k} A^k_{-j}\{(z-a_k)^{-\nu-j} - (\zeta - a_k)^{-\nu-j}\}, \tag{16}$$

where ζ is any point distinct from the poles $\zeta \neq a_k$, with $k = 1, \cdots, M$.

To extend the theorem to the case when the differintegration $D^\nu F/Dz^\nu$ is a meromorphic function with an infinite $M = \infty$ number of poles, we must first consider their location in the complex z-plane, as illustrated in Fig. 2. (i) The poles a_k with $k = 1, \cdots, M$, can have no accumulation point (because the latter would be an essential singularity, contradicting that $D^\nu F/Dz^\nu$ be meromorphic); (ii) a compact region, e.g., a circle $|z| < R$ of center at the origin and radius R, can contain only a finite number of poles (otherwise, if the number of poles were infinite, at least one accumulation point would exist); (iii) if the poles are ordered by modulus $|a_k| \leqq |a_{k+1}|$, they must tend to infinity $|a_k| \to \infty$ as $k \to \infty$ (otherwise if $|a_k| < R$ for all $k = 1, \cdots, \infty$, a circle of finite radius would contain an infinite number of poles); (iv) a sequence of regions

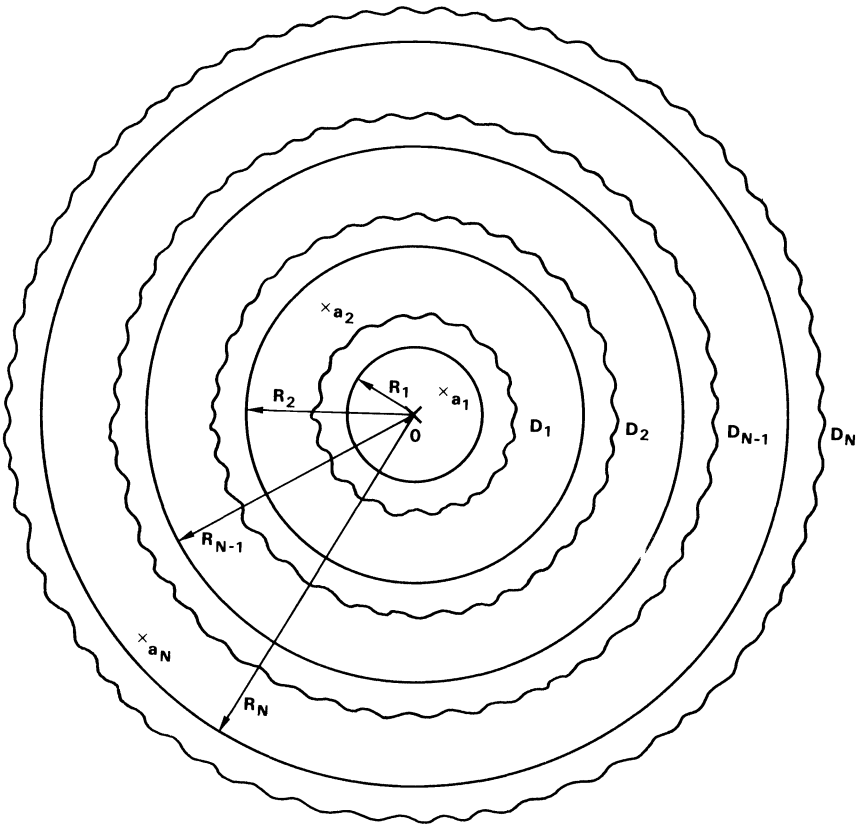


FIG. 2. *If the differintegration of a function is meromorphic, it has an infinity of poles a_k, which when ordered by nondecreasing modulus $|a_{k+1}| \geqq |a_k|$ tend to infinity as $k \to \infty$, allowing the construction of a sequence of regions D_k, each containing the first k poles $a_1, \cdots, a_k$, and all of the preceding regions $D_1, D_2, \ldots, D_{k-1}$, with boundary ∂D_k lying outside a circle of center at the origin, and radius $R_k \to \infty$ growing without bound as $k \to \infty$.*

D_k can be constructed, such that D_k contains only the poles $a_1, \cdots, a_k$ and in its interior can be drawn a circle of radius $R_k \geqq |a_{k-1}|$, such that $R_k \to \infty$ as $k \to \infty$. These remarks show that to extend (16) to meromorphic functions, the following modifications are needed: (i) the second term on the right-hand side should be omitted, since the point at infinity of a meromorphic function is not a pole, but rather a "point of accumulation" of poles; (ii) the sum over the number of poles $M = \infty$, in the last term on the right-hand side of (16) becomes a series, provided that its convergence be assumed by the asymptotic behavior of the differintegration $D^\nu F/Dz^\nu$ as $z \to \infty$.

It will turn out that, in the present case, a sufficient condition for the convergence stated in (ii), is the same as for the original Mittag-Leffler theorem, namely [24], that the differintegration $D^\nu F/Dz^\nu$ be bounded on the boundaries ∂D_k of the sequence of regions D_k tending to infiinity in Fig. 2. This leads to the generalized Mittag-Leffler theorem, for the differintegration of meromorphic functions.

THEOREM 6 (series of fractions). *The differintegration $D^\nu F/Dz^\nu$, with complex order ν of a meromorphic function $F(z)$, with an infinity of poles a_k of orders b_k, with $k = 1, \cdots, \infty$, with coefficients A^k_{-j} with $j = 1, \cdots, b_k$ of the principal parts* (13), *can be represented by a series of fractions*:

$$\frac{D^\nu F}{Dz^\nu} = \frac{D^\nu F}{D\zeta^\nu} + \sum_{k=1}^{\infty} \sum_{j=1}^{b_k} A^k_{-j} \{(z-a_k)^{-\nu-j} - (\zeta - a_k)^{-\nu-j}\}, \tag{17}$$

where $\zeta \neq a_k$ for $k = 1, \cdots, \infty$, is any regular point. Assuming that the differintegration $D^\nu F/Dz^\nu$ is bounded, as $k \to \infty$ on a sequence of loops ∂D_k, each containing the first N poles $a_1, \cdots, a_k$ (Fig. 2), the series (17) *converges*: (i) *simply outside the poles, i.e., for $|z - a_k| > 0$ for all $k = 1, \cdots, \infty$*; (ii) *uniformly outside a closed neighborhood of the poles, viz., for $|z - a_k| \geqq \delta_k > 0$ for some δ_k, with $k = 1, \cdots, \infty$.*

Remark 9. The existence of a pole at infinity, as in the second term on the right-hand side of (16), is incompatible with the boundedness condition on the sequence of loops ∂D_N, which assures the convergence of (17).

Remark 10. The asymptotic condition (2) on the function $F(z)$, even if it could be restated in terms of the differintegration $D^\nu F/Dz^\nu$, would not satisfy the requirement for convergence of (17); the latter is an isotropic condition, applying as $|z| \to \infty$ for all directions $0 \leqq \arg(z) < 2\pi$, whereas the asymptotic condition (2) is restricted to a narrow sector about the branchline.

Remark 11. The main differences between Theorems 5 and 6 are that: (i) the functions

$$G_M(z) \equiv \frac{D^\nu F}{Dz^\nu} - \sum_{k=1}^{M} A^k_{-j} (z - a_k)^{-\nu-j}, \tag{18}$$

are analytic in the compact region D_M, but not in $D_{M+1}, \cdots$, and thus do not reduce to constants; (ii) the Hankel path in Fig. 1 lies partly outside the region D_N in Fig. 2, so that a loop integral for (18) uses a path distinct from Hankel's.

Remark 11 points to the two difficulties (i) and (ii) that must be resolved to prove the convergence of (17) and, in fact, lead to the restrictions in Theorem 6, which appear in addition to those in Theorem 5.

Proof. Since the function (18) is analytic in the region D_M, we may use the generalized Cauchy integral for its differintegration:

$$\frac{D^\nu \{G_M(z)\}}{Dz^\nu} = \{\Gamma(1+\nu)/2\pi i\} \int_{\partial D_M} (\eta - z)^{-\nu-1} G_M(\eta)\, d\eta + \delta_M(z), \tag{19}$$

where the last term on the right-hand side of (19) accounts for the part of the Hankel path (Fig. 1) outside the loop ∂D_M (Fig. 2) and gives a negligible contribution $\delta_M(z) \to 0$ as $M \to \infty$, because of the asymptotic condition (2). The remainder of the series (17) after M terms, is given:

$$\begin{aligned}\Delta_M(z,\zeta) &\equiv \sum_{k=M+1}^{\infty} \sum_{j=1}^{b_k} A_{-j}^{k}\{(z-a_k)^{-\nu-j} - (\zeta-a_k)^{-\nu-j}\} \\ &= G_M(z) - G_M(\zeta),\end{aligned} \tag{20}$$

where (18) was used. Substituting (19) with $\nu = 0$ into (20), we obtain the following integral:

$$\begin{aligned}&\Delta_M(z,\zeta) - \delta_M(z) + \delta_M(\zeta) \\ &= (2\pi i)^{-1} \int_{\partial D_M} G_M(\eta)\eta^{-1}\{(1-z/\eta)^{-1} \\ &\quad -(1-\zeta/\eta)^{-1}\}\, d\eta \\ &= (2\pi i)^{-1}(z-\zeta) \int_{|\eta|=R_M} \eta^{-2}\{1+O(\eta^{-1})\}G_M(\eta)\, d\eta,\end{aligned} \tag{21}$$

which has an upper bound:

$$\begin{aligned}|\Delta_M(z,\zeta) - \delta_M(z) + \delta_M(\zeta)| &\leqq (B|2\pi)\cdot|z-\zeta| \\ &\quad \cdot 2\pi R_M \cdot R_M^{-2}\{1+O(R_M^{-1})\},\end{aligned} \tag{22}$$

because: (i) the differintegration $D^\nu F/Dz^\nu$, and hence (18), is bounded by B on the sequence of loops ∂D_M; (ii) the points z, ζ are fixed points and hence $|\zeta - z|$ is bounded; (iii) the other bounds for terms in (21) show that (22) is $O(R_M^{-1})$, and since $R_M \to \infty$ as $M \to \infty$, the expression (22) vanishes. Also, $\delta_M(z)$, $\delta_M(\zeta) \to 0$ as $M \to \infty$, implying that the remainder of the series (17) vanishes $\Delta_M(z,\zeta) \to 0$ as $M \to \infty$, proving its simple convergence. The convergence is improved from simple to uniform for points at a finite distance $|\zeta - a_k| \geqq \delta_k > 0$ from all $k = 1, \cdots, \infty$ the poles a_k. □

4. Rules of differintegration of products and implicit functions. The prototype of series expansion for a differintegration is the extended Taylor series (3) from which the theory may be developed in two directions: (i) by allowing one or more singularities of the function $F(z)$, we are led to the extended Laurent series (4)-(6), the principal parts in the neighborhood of singularities (8), (9), and the generalized Mittag-Leffler theorem (17) developed in §§ 2-3; (ii) in §§ 4-5, we consider the consequences of replacing the independent variable z by an auxiliary analytic function $f(z)$ in the extended Taylor series (3), leading to an extended Lagrange series:

$$\frac{D^\nu\{F(z)\}}{D\{f(z)\}^\nu} = \sum_{k=0}^{\infty} \{f(z)\}^k (k!)^{-1} \frac{D^{\nu+k}\{F(\zeta)\}}{D\{f(\zeta)\}^{\nu+k}}, \tag{23}$$

provided that the implicit differintegration appearing in the last factor of (23) can be evaluated. The extended Lagrange series (23), expressing the differintegration of an analytic function as an ascending power series of a suitable auxiliary analytic function $f(z)$, can be further generalized [3], [6] to the extended Teixeira series; the latter allows the dependent variable $F(z)$ to have an isolated singularity at $z = \zeta$, in which case the power series expansion involves descending as well as ascending powers of

the independent variable $f(z)$, e.g., the particular case $f(z)=z$ is the extended Laurent series (4)–(6). Our present aim is not to obtain the most general power series expansion, as that question is addressed elsewhere [19], [21], [4], [6]. Thus, we leave these remarks in passing and return to the rule of implicit differintegration, which is of use in connection with problems other than the extended Lagrange series (23).

If in Definition 1 of differintegration (1) the independent variable z is replaced by an analytic function $f(z)$, we obtain the integral

$$\frac{D^{\nu}\{F(z)\}}{D\{f(z)\}^{\nu}} \equiv \{\Gamma(1+\nu)/2\pi i\} \int_{\infty \exp\{i \arg (f(z))\}}^{(f(z)+)} \{f(\zeta)-f(z)\}^{-\nu-1} F(\zeta) f'(\zeta)\, d\zeta, \tag{24}$$

the evaluation of which leads to the rule of implicit differintegration.

PROPERTY 4. The implicit differintegration (24), with complex order ν of an analytic function $F(z)$, with regard to an auxiliary analytic function $f(z)$, at a point $\zeta = z$ that is not a zero of the latter $f'(z) \neq 0$, is given by

$$\frac{D^{\nu}\{F(z)\}}{D\{f(z)\}^{\nu}} = \lim_{\zeta \to z} \left(\frac{\partial}{\partial \zeta}\right)^{\nu} \{F(\zeta)\{(\zeta - z)/(f(\zeta)-f(z))\}^{\nu+1} f'(\zeta)\}. \tag{25}$$

Remark 12. The assumption that the auxiliary analytic function $f(\zeta)$ is not zero at $\zeta = z$ implies that the equation

$$f(\zeta)-f(z) = (\zeta - z) H(\zeta, z), \tag{26a}$$

$$H(z, z) \neq 0, \tag{26b}$$

has a single root $\zeta = z$, and thus the function

$$G(\zeta, z) \equiv F(\zeta)\{(f(\zeta)-f(z))/(\zeta - z)\}^{-\nu-1} f'(\zeta) \tag{27}$$

is also analytic at $\zeta = z$.

Proof. The function (27) may be substituted into (24), which gives

$$\frac{D^{\nu}\{F(z)\}}{D\{f(z)\}^{\nu}} = \{\Gamma(1+\nu)/2\pi i\} \int G(\zeta, z)(\zeta - z)^{-\nu-1}\, d\zeta = \lim_{\zeta \to z} \frac{\partial^{\nu}\{G(\zeta, z)\}}{\partial \zeta^{\nu}}, \tag{28}$$

allowing its evaluation by (1). The expression (28) with (27) leads to (25). □

We may apply Property 4 to the evaluation of the coefficients of the extended Lagrange series (23), viz., the following theorem.

THEOREM 7 (extended Lagrange series). *The differintegration* (24) *with complex order* ν *of an analytic function* $F(z)$, *with regard to an auxiliary analytic function* $f(z)$, *without zeros* $f'(z) \neq 0$, *can be expanded in an ascending power series of the latter* (23), *with coefficients given by* (25). *The series converges*: (i) *absolutely in the open region*:

$$D - \partial D \equiv \{z: |f(z)| < R\}; \tag{29}$$

(ii) *uniformly in the closed subregion*:

$$0 < \varepsilon < R\text{: } D_{\varepsilon} \equiv \{z: |f(z)| \leqq R - \varepsilon\} \subset D - \partial D, \tag{30}$$

where R *is the largest positive real number such that the region* D *excludes all singularities of* $F(z)$.

Proof. The only results not proved before concern the conditions of convergence, which are similar to those for the extended Taylor series in Theorem 3, replacing $f(z)=z-\zeta$, as suggested by a comparison of (23) and (3). □

To develop the rule of implicit differintegration (Property 4) further, we need the extended Leibnitz rule, in one of the forms [20], [22], [2], [4], that appear in the literature, viz., Property 5.

PROPERTY 5. The differintegration (1) of complex order ν of the product of two analytic functions $F(z)E(z)$, such that $F(z)$ and $F(z)E(z)$ satisfy the asymptotic condition (2), is given by the infinite chain rule:

$$\frac{D^{\nu}\{F(z)E(z)\}}{Dz^{\nu}}=\sum_{k=0}^{\infty}\binom{\nu}{k}E^{(k)}(z)\frac{D^{\nu+k}F}{Dz^{\nu+k}}, \tag{31}$$

$$\binom{\nu}{k}\equiv\Gamma(1+\nu)/\{k!\Gamma(1+\nu-k)\}. \tag{32}$$

Proof. A direct proof of the extended Leibnitz rule in the form (31) can be found elsewhere [5].

To apply the infinite chain rule (31) to (25), we split the term in curly brackets into two factors, namely, $F(\zeta)$ and $E(z,\zeta)$:

$$\frac{d^{\nu}\{F(z)\}}{d\{f(z)\}^{\nu}}=\lim_{\zeta\to z}\left(\frac{\partial}{\partial\zeta}\right)^{\nu}\{F(\zeta)E(z,\zeta)\}, \tag{33}$$

$$E(z,\zeta)\equiv f'(\zeta)\{(\zeta-z)/(f(\zeta)-f(z))\}^{\nu+1}, \tag{34}$$

to obtain Property 6.

PROPERTY 6. The differintegration (24), with complex order ν of an analytic function $F(z)$, with regard to an auxiliary analytic function $f(z)$, at a point z that is not a zero of the latter $f'(z)\neq 0$, is given by

$$\begin{aligned}\frac{D^{\nu}\{F(z)\}}{D\{f(z)\}^{\nu}}&=\sum_{k=0}^{\infty}\binom{\nu}{k}\frac{D^{k}F}{Dz^{k}}\frac{D^{\nu-k}E}{Dz^{\nu-k}}\\&=\sum_{k=0}^{\infty}\binom{\nu}{k}\frac{D^{k}E}{Dz^{k}}\frac{D^{\nu-k}F}{Dz^{\nu-k}},\end{aligned} \tag{35}$$

$$\frac{D^{\mu}E}{Dz^{\mu}}\equiv\lim_{\zeta\to z}\left(\frac{\partial}{\partial\zeta}\right)^{\mu}f'(\zeta)\{(\zeta-z)/(f(\zeta)-f(z))\}^{\nu+1}, \tag{36}$$

where $F(z)$, and $F(z)E(z,\zeta)$ are assumed (see (34)) to satisfy the asymptotic condition (3).

Proof. Substitute (31) into (33) to obtain (35), with (36) given by (34). □

The application of the rule (35) of implicit differintegration is facilitated by the following remark.

Remark 13. The series (35) terminates at $k=n$: (i) for $\nu=n$ a positive integer, i.e., an ordinary derivative of order n, or (ii) if one of the functions $f(z)$, $E(z,\zeta)$ has derivatives vanishing beyond the order n, i.e., is a polynomial.

The latter case (ii) can be illustrated by the implicit differintegration of a power $F(z)=z^{m}$, with a positive integral exponent m:

$$\frac{d^{\nu}(z^{m})}{d\{f(z)\}^{\nu}}=\sum_{k=0}^{m}\binom{\nu}{k}\{m!/(m-k)!\}z^{m-k}\frac{d^{\nu-k}E}{dz^{\nu-k}}; \tag{37}$$

in (37) we have used Riemann d^{ν} instead of Liouville D^{ν} differintegrations, since Properties 4–6 hold for the former [1], [2], as well as for the latter. Of the two series

(35), one may be simpler than the other, e.g., for the implicit differintegration of the logarithm:

$$\frac{d^\nu(\log z)}{d\{f(z)\}^\nu}=\log z\frac{d^\nu E}{dz^\nu}+\sum_{k=1}^{\infty}\{(-)^k/k\}\{\Gamma(1+\nu)/\Gamma(1+\nu-k)\}z^{-k}\frac{d^{\nu-k}E}{dz^{\nu-k}}. \tag{38}$$

Both in (37) and (38) the coefficients (36) still need to be calculated; an example is the case of implicit differintegration with regard to the square $f(z)=z^2$, away from the origin $f'(z)=2z\neq 0$, viz.,

$$\begin{aligned}\frac{d^{\nu-k}E}{dz^{\nu-k}}&=\lim_{\zeta\to z}\left(\frac{\partial}{\partial\zeta}\right)^{\nu-k}\{2\zeta(\zeta+z)^{-\nu-1}\}\\ &=\lim_{\zeta\to z}2\left\{\zeta\left(\frac{\partial}{\partial\zeta}\right)^{\nu-k}+(\nu-k)\left(\frac{\partial}{\partial\zeta}\right)^{\nu-k-1}\right\}(\zeta+z)^{-\nu-1}\\ &=\{\Gamma(1-\nu)/\Gamma(1+k-2\nu)\}(2z)^{k-2\nu}.\end{aligned} \tag{39}$$

Substituting (39) into (37) and (38), we obtain

$$\frac{d^\nu(z^m)}{d(z^2)^\nu}=\sum_{k=0}^{m}\binom{m}{k}C_{\nu,k}2^{k-2\nu}z^{m-2\nu}, \tag{40}$$

$$\frac{d^\nu(\log z)}{d(z^2)^\nu}=\log z\{\Gamma(1-\nu)/\Gamma(1-2\nu)\}(2z)^{-2\nu}+\sum_{k=1}^{\infty}\left\{\frac{(-)^k}{k}\right\}2^{k-2\nu}z^{-2\nu}C_{\nu,k}, \tag{41}$$

$$C_{\nu,k}\equiv\Gamma(1+\nu)\Gamma(1-\nu)/\Gamma(1+\nu-k)\Gamma(1+k-2\nu), \tag{42}$$

as worked-out examples of the implicit differintegration rule.

5. Discussion. The differintegration applied most often to special functions [11], [10], [2], [4] is the Riemann type, since the hypergeometric function and several of its particular cases involve branchpoints. The Liouville differintegration [17] is not suitable for these functions, but rather for those defined from analytic functions; an example is the Hermite function of complex order ν and variable z, which is defined [5] by the Liouville differintegration of the Gaussian function:

$$H_\nu(z)\equiv e^{i\pi\nu}\,e^{z^2}\frac{D^\nu(e^{-z^2})}{Dz^\nu}. \tag{43}$$

Substituting ν by $\nu+1$, and using the chain rule (31):

$$\begin{aligned}e^{i\pi(\nu+1)}\,e^{z^2}\left(\frac{D}{Dz}\right)^{\nu+1}e^{-z^2}&=e^{i\pi\nu}\,e^{z^2}\left(\frac{D}{Dz}\right)^{\nu}\{2z\,e^{-z^2}\}\\ &=e^{i\pi\nu}\,e^{z^2}\left\{2z\left(\frac{D}{Dz}\right)^{\nu}+2\nu\left(\frac{D}{Dz}\right)^{\nu-1}\right\}e^{-z^2},\end{aligned} \tag{44}$$

we obtain the recurrence formula

$$H_{\nu+1}(z)=2zH_\nu(z)-2\nu H_{\nu-1}(z). \tag{45}$$

Differentiating (43) with regard to z:

$$\left(\frac{D}{Dz}\right)e^{i\pi\nu}\,e^{z^2}\left(\frac{D}{Dz}\right)^{\nu}e^{-z^2}=e^{i\pi\nu}\,e^{z^2}\left\{2z\left(\frac{D}{Dz}\right)^{\nu}+\left(\frac{D}{Dz}\right)^{\nu+1}\right\}e^{-z^2}, \tag{46}$$

we obtain the derivation formula

$$\frac{D\{H_\nu(z)\}}{Dz} = 2zH_\nu(z) - H_{\nu+1}(z) = 2\nu H_{\nu-1}(z), \tag{47}$$

which was simplified using the recurrence formula (45).

Both the recurrence (45) and derivation (47) formulas hold for all complex values of ν and lead to

$$\begin{aligned}\frac{D^2\{H_\nu(z)\}}{Dz^2} &= \frac{D\{2zH_\nu(z)\}}{Dz} - \frac{D\{H_{\nu+1}(z)\}}{Dz} \\ &= 2H_\nu(z) + 2z\frac{D\{H_\nu(z)\}}{Dz} - 2(\nu+1)H_\nu(z),\end{aligned} \tag{48}$$

which is the differential equation satisfied by Hermite functions,

$$\left\{\frac{D^2}{Dz^2} - 2z\frac{D}{Dz} + 2\nu\right\} H_\nu(z) = 0. \tag{49}$$

Substituting (1) into (43), we obtain the integral representation of Hermite functions:

$$H_\nu(z) = \{\Gamma(1+\nu)/2\pi i\}\, e^{z^2} \int_{\infty \exp\{i \arg(z)\}}^{(z+)} (z-\zeta)^{-\nu-1}\, e^{-\zeta^2}\, d\zeta; \tag{50}$$

the change of variable $\eta = z - \zeta$ can be performed to move the branchpoint $\zeta = z$ of the integrand to the origin $\eta = 0$ of the η-plane:

$$H_\nu(z) = \{\Gamma(1+\nu)/2\pi i\} \int_{\infty \exp(i0)}^{(0+)} \eta^{-\nu-1}\, e^{-\eta^2+2\eta z}\, d\eta; \tag{51}$$

this can be interpreted again using (1) to show that the Hermite function is specified by the limit:

$$H_\nu(z) = \lim_{\eta \to 0} D^\nu\{\exp(-\eta^2 + 2\eta z)\}/D\eta^\nu. \tag{52}$$

Using the chain rule (31) once more:

$$\frac{D^\nu\{e^{-\eta^2}\, e^{2\eta z}\}}{D\eta^\nu} = e^{2\eta z} \sum_{k=0}^{\infty} \binom{\nu}{k} (2z)^k \frac{D^{\nu-k}(e^{-\eta^2})}{D\eta^{\nu-k}}, \tag{53}$$

in (52), we obtain the expansion of the Hermite function in power series:

$$H_\nu(z) = e^{-i\pi\nu} \sum_{k=0}^{\infty} \binom{\nu}{k} (-2z)^k H_{\nu-k}(0), \tag{54}$$

with coefficients involving its value at the origin, e.g., for $\nu = \eta$ a positive integer:

$$H_n(z) = (-)^n \sum_{k=0}^{n} \binom{n}{k} (-2z)^k H_{n-k}(0), \tag{55}$$

we have an identity for Hermite [9] polynomials.

We conclude the examples of differintegrations applied to Hermite functions with the deduction of an addition formula. The latter is derived from the definition (43):

$$H_\nu(z) = e^{i\pi\nu}\, e^{z^2} (D/Dz)^\nu \{e^{-az^2}\, e^{-(1-a)z^2}\}, \tag{56}$$

by applying the chain rule (31):

$$\left(\frac{D}{Dz}\right)^{\nu}\{e^{-az^2}\,e^{-(1-a)z^2}\}=\sum_{k=0}^{\infty}\binom{\nu}{k}\left(\frac{D}{Dz}\right)^{k}e^{-az^2}\left(\frac{D}{Dz}\right)^{\nu-k}e^{-(1-a)z^2}. \tag{57}$$

Introducing the new variables:

$$x\equiv z\sqrt{a}, \tag{58a}$$

$$y\equiv z\sqrt{1-a}, \tag{58b}$$

the factors in (57) become, respectively,

$$\left(\frac{D}{Dz}\right)^{k}e^{-az^2}=a^{k/2}\left(\frac{D}{Dx}\right)^{k}e^{-x^2}, \tag{59a}$$

$$\left(\frac{D}{Dz}\right)^{\nu-k}e^{-(1-a)z^2}=(1-a)^{(\nu-k)/2}\left(\frac{D}{Dy}\right)^{\nu-k}e^{-y^2}. \tag{59b}$$

Substitution of (59a), (59b) into (57) and (56) gives

$$\begin{aligned}H_{\nu}(z)=\sum_{k=0}^{\infty}\binom{\nu}{k}e^{i\pi k}\,e^{x^2}\left(\frac{D}{Dx}\right)^{\nu}e^{-x^2}a^{k/2}\\ \cdot(1-a)^{(\nu-k)/2}\,e^{i\pi(\nu-k)}\,e^{y^2}\left(\frac{D}{Dy}\right)^{\nu-k}e^{-y^2},\end{aligned} \tag{60}$$

where we have used (61a):

$$x^2+y^2=z^2, \tag{61a}$$

$$a=1/(1+y^2/x^2), \tag{61b}$$

which follows from (58a), (58b) together with (61b). Recalling (43), we may interpret (60), with (61a), (61b), as Property 7.

PROPERTY 7. The Hermite functions satisfy the addition theorem:

$$\begin{aligned}H_{\nu}(\sqrt{x^2+y^2})=\sum_{k=0}^{\infty}\binom{\nu}{k}(1+y^2/x^2)^{-k/2},\\ H_k(x)(1+x^2/y^2)^{-(\nu-k)/2}H_{\nu-k}(y),\end{aligned} \tag{62}$$

which terminates only at $k=n$, a positive integer, only for Hermite polynomials H_n.

REFERENCES

[1] L. M. B. C. CAMPOS, *On a concept of derivative of complex order, with application to special functions*, IMA J. Appl. Math., 33 (1984), pp. 109-133.

[2] ———, *On rules of derivation with complex order of analytic and branched functions*, Portugal. Math., 43 (1985), pp. 347-376.

[3a] ———, *On extensions of Laurent's theorem in the fractional calculus, with application to the generation of higher transcendental functions*, Mat. Vesnik, 38 (1986), pp. 375-390.

[3b] ———, *Erratum*, 40 (1988), p. 85.

[4] ———, *On a systematic approach to some properties of special functions*, IMA J. Appl. Math., 36(1986), pp. 191-206.

[5] ———, *On the branch-point operator and the annihilation of differintegrations*, SIAM J. Math. Anal., this issue, pp. 439-453.

[6] ———, *On generalizations of the theorems of Taylor, Lagrange, Laurent and Teixeira*, 1987, unpublished.

[7] A. ERDELYI, ed., *Higher Transcendental Functions*, 3 vols., McGraw-Hill, New York, 1953.

[8] H. HANKEL, *Die Euler'schen integrale bei unbeschrankter Variabilitat des Argumentes*, Z. Mat. Phys., 9 (1864), pp. 7-11.

[9] C. HERMITE, *Sur un nouveau developpement en série de fonctions*, Compt. Rend. Acad. Sci. Paris, 58 (1884), pp. 83-100 and 266-273; Oeuvres 2 (1884), pp. 293-302.

[10] J. L. LAVOIE, T. J. OSLER, AND R. TREMBLAY, *Fractional derivatives and special functions*, SIAM Rev., 18 (1976), pp. 240-268.

[11] ———, *Fundamental properties of fractional derivatives via Pochhammer integrals*, in Fractional Calculus and Applications, B. Ross, ed., Springer-Verlag, Berlin, New York, 1974, pp. 323-356.

[12] A. V. LETNIKOV, *Theory of differentiation of fractional order*, Mat. Sb., 3 (1868), pp. 1-60.

[13] J. LIOUVILLE, *Mémoire sur le calcul des differentielles à indices quelconques*, Journ. Éc. Polyt., 13 (1832), pp. 71-162.

[14] A. C. MCBRIDE, *Fractional Calculus and Integral Transforms of Generalized Functions*, Pitman, Boston, 1979.

[15] A. C. MCBRIDE AND G. F. ROACH, eds., *Fractional Calculus*, Pitman, Boston, 1986.

[16] P. A. NEKRASSOV, *Generalized differentiation*, Mat. Sb., 14 (1888), pp. 45-168.

[17] K. NISHIMOTO, *Fractional Calculus*, 2 vols., Descartes Press, Koryama, Japan, 1984.

[18] K. B. OLDHAM AND J. SPANIER, *Fractional Calculus*, Academic Press, New York, 1974.

[19] T. J. OSLER, *The fractional derivative of a composite function*, SIAM J. Math. Anal., 1 (1970), pp. 288-293.

[20] ———, *Leibnitz rule for fractional derivatives generalized and an application to infinite series*, SIAM J. Appl. Math., 18 (1970), pp. 658-674.

[21] ———, *Taylor's series generalized for fractional derivatives and applications*, SIAM J. Math. Anal., 2 (1971), pp. 37-48.

[22] ———, *A further extension of the Leibnitz rule to fractional derivatives and its relation to Parseval's formula*, SIAM J. Math. Anal., 3 (1972), pp. 1-16.

[23] B. ROSS, ed., *Fractional Calculus and Applications*, Springer-Verlag, Berlin, New York, 1974.

[24] E. T. WHITTAKER AND G. N. WATSON, *Course of Modern Analysis*, Cambridge University Press, Cambridge, 1902; 6th ed., 1927.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 468-478, March 1989

014

EXISTENCE THEOREMS FOR BOUNDARY VALUE PROBLEMS FOR *n*th-ORDER NONLINEAR DIFFERENCE EQUATIONS*

JOHNNY HENDERSON†

Abstract. For the nth-order nonlinear difference equation $u(m+n)=f(m, u(m), u(m+1), \cdots, u(m+n-1))$, where $f:[a, +\infty)\times \mathbf{R}^n \to \mathbf{R}$ is continuous, and the equation $u_n = f(m, u_0, \cdots, u_{n-1})$ can be solved for u_0 as a continuous function of $u_1, \cdots, u_n$ for each $m \in [a, +\infty)$, it is shown that the uniqueness of solutions implies the existence of solutions for conjugate boundary value problems on $[a, +\infty)$. Shooting methods are used in conjunction with an induction.

Key words. nonlinear difference equation, boundary value problem, uniqueness implies existence

AMS(MOS) subject classifications. 34B10, 34B15, 39A10, 39A12

1. Introduction. For $a \in \mathbf{R}$, let the interval $[a, +\infty)=\{a, a+1, a+2, \cdots\}$, and if $b=a+m$, for some $m \in \mathbf{N}$, let the closed interval $[a, b]=\{a, a+1, \cdots, b\}$, and let the intervals $[a, b)$, $(a, b]$, and (a, b) denote the analogous discrete sets. In this paper, we will be concerned with uniqueness of solutions implying the existence of solutions of certain boundary value problems for the nth-order nonlinear difference equation

$$u(m+n)=f(m, u(m), u(m+1), \cdots, u(m+n-1)), \qquad n \geqq 2, \tag{1}$$

where

(A) $f:[a, +\infty)\times \mathbf{R}^n \to \mathbf{R}$ is continuous and the equation $u_n = f(m, u_0, \cdots, u_{n-1})$ can be solved for u_0 as a continuous function of $u_1, \cdots, u_n$ for each $m \in [a, +\infty)$.

We remark here that (A) implies (1) is an nth-order difference equation on any subinterval of $[a, +\infty)$, that solutions of initial value problems for (1) are unique and exist on $[a, +\infty)$, and that solutions of (1) depend continuously on initial conditions.

Recently, much research activity has focused on the existence of solutions of boundary value problems for finite difference equations. Much of this activity has been motivated by Hartman's landmark paper [18] and has been devoted to analogues of results for boundary value problems for ordinary differential equations. In particular, Ahlbrandt and Hooker [7]-[8], Eloe [11]-[13], Hankerson [14], and Peterson [25]-[28] have published a number of results concerning the disconjugacy or disfocality of linear difference equations, whereas other papers by Ahlbrandt and Hooker [6], Hankerson and Peterson [15]-[16], Hooker et al. [19]-[22], Patula [24], and Smith and Taylor [31] have dealt with oscillation and nonoscillation of linear difference equations. Moreover, in papers by Agarwal [1]-[4], Eloe [9]-[10], Hankerson [14], Peterson [29], and Rodriguez [30], questions have been addressed dealing with boundary value problems for the nonlinear difference equation (1).

In this work, the types of boundary value problems for (1), for which we address the question of uniqueness of solutions implying the existence of solutions, are analogous in some sense to those known as conjugate problems for ordinary differential equations.

DEFINITION. Given $m_1 \in [a, +\infty)$ and $m_2, \cdots, m_n \in \mathbf{N}$, let $s_1, \cdots, s_n \in [a, +\infty)$ be defined by $s_1 = m_1$ and $s_i = s_{i-1} + m_i$, $2 \leqq i \leqq n$. A boundary value problem for (1) satisfying the conditions

$$u(s_i) = y_i, \qquad 1 \leqq i \leqq n, \tag{2}$$

* Received by the editors July 20, 1987; accepted for publication (in revised form) June 1, 1988.
† Department of Algebra, Combinatorics and Analysis, Auburn University, Auburn, Alabama 36849.

where $y_i \in \mathbf{R}$, $1 \leqq i \leqq n$, will be called an $(m_1, \cdots, m_n)$ *conjugate boundary value problem for* (1). (In the case when $m_i = 1$, $2 \leqq i \leqq n$, the problem becomes an initial value problem for (1).)

For $(m_1, \cdots, m_n)$ conjugate boundary value problems for (1), our main result is motivated by a theorem for conjugate boundary value problems for ordinary differential equations proved by Hartman [17] and Klaasen [23].

We will use some of the terminology introduced by Hartman [18] and employed in many of the above-cited papers dealing with difference equations. For a function $u : [a, +\infty) \to \mathbf{R}$, Hartman defined $m_0 \in [a, +\infty)$, in the case that $m_0 = a$, to be a *node* of u if $u(a) = 0$, and $m_0 > a$ to be a *node* of u if $u(m_0) = 0$ or $u(m_0 - 1)u(m_0) < 0$. Moreover, Hartman defined $m_0 = a$ to be a *generalized zero* of u if $u(a) = 0$, and $m_0 > a$ to be a *generalized zero* of u if $u(m_0) = 0$ or there is an integer $j \geqq 1$ such that $(-1)^j u(m_0 - j)u(m_0) > 0$ and if $j > 1$, $u(m_0 - j + 1) = \cdots = u(m_0 - 1) = 0$. (Note that if m_0 is a node of u, then m_0 is a generalized zero of u.)

In view of this terminology, our uniqueness assumption on $(m_1, \cdots, m_n)$ conjugate boundary value problems for (1) takes the following form:

(B) Given $m_1 \in [a, +\infty)$ and $m_2, \cdots, m_n \in \mathbf{N}$, if $s_1 = m_1$ and $s_i = s_{i-1} + m_i$, $2 \leqq i \leqq n$, and if $u(m)$ and $v(m)$ are solutions of (1) such that $u(s_1) = v(s_1)$ and $u(m) - v(m)$ has a generalized zero at s_i, $2 \leqq i \leqq n$, then it follows that $u(m) = v(m)$ on $[s_1, s_n]$ (hence on $[a, +\infty)$).

In § 2, we will state for convenience and reference some theorems concerning continuous dependence of solutions of (1) on initial and boundary conditions. Then, in § 3, we prove that under assumptions (A) and (B), each $(m_1, \cdots, m_n)$ conjugate boundary value problem for (1) has a unique solution on $[a, +\infty)$. The proof employs shooting methods in conjunction with an induction on the indices $m_2, \cdots, m_n$.

2. Continuous dependence. In this section, we will state standard theorems concerning the continuous dependence of solutions of initial value problems and $(m_1, \cdots, m_n)$ conjugate boundary value problems for (1). We state these results for convenience and reference, and hence will eliminate repeating them later.

Our first theorem follows immediately from condition (A).

THEOREM 1. *Assume that condition* (A) *is satisfied. If there exist a sequence* $\{y_k(m)\}$ *of solutions of* (1), *an interval* $[m_0, m_0 + n - 1] \subset [a, +\infty)$, *and an* $M > 0$ *such that* $|y_k(m)| \leqq M$, *for all* $m \in [m_0, m_0 + n - 1]$, *for all* $k \in \mathbf{N}$, *then there exists a subsequence* $\{y_{k_j}(m)\}$ *that converges pointwise on* $[a, +\infty)$ *to a solution of* (1).

In turn it follows that if (B) is also assumed, then the continuous dependence of solutions on initial conditions, coupled with an application of the Brouwer Theorem on the invariance of domain, imply that solutions of conjugate problems depend continuously on boundary conditions; for a typical argument with difference equations, see [14].

THEOREM 2. *Assume that with respect to* (1), *conditions* (A) *and* (B) *are satisfied. Given a solution* $u(m)$ *of* (1) *on* $[a, +\infty)$, *points* $s_1 < s_2 < \cdots < s_n$ *belonging to* $[a, +\infty)$, *an interval* $[s_1, b]$, *where* $b \geqq s_n$, *and* $\varepsilon > 0$, *there exists* $\delta(\varepsilon, [s_1, b]) > 0$ *such that, if* $|u(s_i) - y_i| < \delta$, $1 \leqq i \leqq n$, *then there exists a solution* $v(m)$ *of* (1) *satisfying* $v(s_i) = y_i$, $1 \leqq i \leqq n$, *and* $|v(m) - u(m)| < \varepsilon$, *for all* $m \in [s_1, b]$.

From Theorem 2, we can state a theorem similar to Theorem 1.

THEOREM 3. *Assume that* (A) *and* (B) *are satisfied and suppose that, given* $m_1 \in [a, +\infty)$ *and for some* $m_2, \cdots, m_n \in N$, *there exist unique solutions of* (1), (2). *Let* s_i, $1 \leqq i \leqq n$, *be corresponding points in* $[a, +\infty)$. *If there exist a sequence* $\{y_k(m)\}$ *of solutions of* (1) *and an* $M > 0$ *such that* $|y_k(s_i)| \leqq M$, $1 \leqq i \leqq n$, *for all* $k \in N$, *then there exists a*

subsequence $\{y_{k_j}(m)\}$ *that converges pointwise on* $[a, +\infty)$. *In particular, for this subsequence, if* $\lim_j y_{k_j}(s_i) = y_i$, $1 \leqq i \leqq n$, *then* $\{y_{k_j}(m)\}$ *converges pointwise on* $[a, +\infty)$ *to the solution of the* $(m_1, \cdots, m_n)$ *conjugate boundary value problem for* (1) *satisfying*

$$y(s_i) = y_i, \qquad 1 \leqq i \leqq n.$$

3. Uniqueness implies existence. In this section, we prove that hypotheses (A) and (B) imply the existence of solutions of $(m_1, \cdots, m_n)$ conjugate boundary value problems for (1). The method of shooting is employed in conjunction with an induction on $m_2, \cdots, m_n$. Shooting methods, frequently paired with comparison results for uniqueness, have been used by Agarwal [4], Hankerson [14], and Peterson [29] in obtaining the existence of solutions of two-point problems for (1). For linear difference equations, results appear in [5] where shooting methods are used for multipoint boundary value problems.

The theorem we present here is analogous to the uniqueness implies existence result that Hartman [17] and Klaasen [23] proved for conjugate boundary value problems for ordinary differential equations.

To better illustrate the inductive pattern used with the shooting method in the proof, we will specifically detail the inductive steps on the indices m_{n-2}, m_{n-1}, m_n. Then, the general inductive step is outlined in the latter part of the proof.

THEOREM 4. *Assume that with respect to* (1), *conditions* (A) *and* (B) *are satisfied. Then given* $m_1 \in [a, +\infty)$ *and* $m_2, \cdots, m_n \in N$, *each* $(m_1, \cdots, m_n)$ *conjugate boundary value problem for* (1) *has a unique solution on* $[a, +\infty)$.

Proof. We remark first that the uniqueness of all such solutions follows from (B). As stated above, the proof is by induction on $m_2, \cdots, m_n$, and throughout the proof, let $y_i \in R$, $1 \leqq i \leqq n$, be given.

To begin, let $m_1 \in [a, +\infty)$ be given, let $m_2 = \cdots = m_n = 1$, and let $s_1 = m_1$ and $s_i = s_{i-1} + m_i$, $2 \leqq i \leqq n$. From the existence of unique solutions of initial value problems for (1), it follows that there exists a unique solution $u(m)$ of (1) on $[a, +\infty)$ satisfying

$$u(s_i) = y_i, \qquad 1 \leqq i \leqq n.$$

Assume now that $m_n > 1$ and that, given $m_1 \in [a, +\infty)$ and $m_2 = \cdots = m_{n-1} = 1$, there exists a unique solution of each $(m_1, 1, \cdots, 1, h)$ conjugate boundary value problem, where $1 \leqq h < m_n$, for (1) on $[a, +\infty)$.

Under this assumption, let $m_1 \in [a, +\infty)$ be given, let $m_2 = \cdots = m_{n-1} = 1$, let $s_1 = m_1$, $s_i = s_{i-1} + m_i$, $2 \leqq i \leqq n$, and let $z_1(m)$ be the solution (given by the induction hypotheses), of the $(m_1, 1, \cdots, 1, m_n - 1)$ conjugate boundary value problem for (1) satisfying

$$z_1(s_i) = y_i, \qquad 1 \leqq i \leqq n-1,$$
$$z_1(s_n - 1) = 0.$$

Now define $S_1 = \{r \in \mathbf{R} \mid$ there is a solution $y(m)$ of (1) satisfying $y(s_i) = z_1(s_i)$, $1 \leqq i \leqq n-1$, and $y(s_n) = r\}$. Since $z_1(s_n) \in S_1$, S_1 is nonempty. Moreover, it follows from Theorem 2 that S_1 is an open subset of $\mathbf{R}$.

We claim that S_1 is also a closed subset of $\mathbf{R}$. Assuming the claim is false, it follows that there exist $r_0 \in \bar{S}_1 \backslash S_1$ and a strictly monotone sequence $\{r_k\} \subset S_1$ such that $\lim_k r_k = r_0$. We may assume without loss of generality that $r_k \uparrow r_0$. For each $k \in \mathbf{N}$, let $y_k(m)$ denote the corresponding solution of (1) satisfying

$$y_k(s_i) = z_1(s_i), \qquad 1 \leqq i \leqq n-1,$$
$$y_k(s_n) = r_k.$$

It follows from (B) that $y_k(m) < y_{k+1}(m)$ on $(s_{n-1}, +\infty)$, for all $k \in \mathbf{N}$. Furthermore, the induction hypothesis implies the existence of unique solutions of $(m_1, 1, \cdots, 1, m_n - 1)$ boundary value problems for (1), which, when coupled with Theorem 3 along with $r_0 \notin S$, implies that $y_k(s_n - 1) \uparrow +\infty$, as $k \to +\infty$. Moreover, by Theorem 1, there exists $m_0 \in (s_n, s_n + n - 1]$ such that $y_k(m_0) \uparrow +\infty$, as $k \to +\infty$.

Now let $u(m)$ denote the solution at the points $s_2, \cdots, s_{n-1}, s_{n-1}+1, s_n$ of the $(m_1+1, 1, \cdots, 1, m_n - 1)$ conjugate boundary value problem for (1) satisfying

$$u(s_i) = z_1(s_i), \qquad 2 \leqq i \leqq n-1,$$
$$u(s_{n-1}+1) = 0,$$
$$u(s_n) = r_0.$$

Since $y_k(s_n - 1) \uparrow +\infty$ and $y_k(m_0) \uparrow +\infty$, whereas $y_k(s_n) = r_k < r_0 = u(s_n)$, for all k, it follows that, for some $K \in \mathbf{N}$, $u(m) - y_K(m)$ has a generalized zero at s_n and also a generalized zero (or zero) at some $n_0 \in (s_n, m_0]$. Furthermore, $u(s_i) - y_K(s_i) = 0$, $2 \leqq i \leqq n-1$, and hence from (B), $u(m) = y_K(m)$ on $[a, +\infty)$, a contradiction.

Hence S_1 is also closed and consequently $S_1 = \mathbf{R}$. Choosing $y_n \in S_1$, it follows that there exists a solution $y(m)$ of (1) satisfying

$$y(s_i) = y_i, \qquad 1 \leqq i \leqq n.$$

In particular, given $m_1 \in [a, +\infty)$, $m_2 = \cdots = m_{n-1} = 1$, and $m_n \geqq 1$, each $(m_1, 1, \cdots, 1, m_n)$ conjugate boundary value problem for (1) has a unique solution on $[a, +\infty)$.

For the next part of the proof we induct on m_{n-1}. For this part, we now assume that $m_{n-1} > 1$ and that given $m_1 \in [a, +\infty)$, $m_2 = \cdots = m_{n-2} = 1$, and $m_n \geqq 1$, there exists a unique solution of each $(m_1, 1, \cdots, 1, l, m_n)$ conjugate boundary value problem, where $1 \leqq l < m_{n-1}$, for (1) on $[a, +\infty)$.

With that assumption, let $m_1 \in [a, +\infty)$ be given, let $m_2 = \cdots = m_{n-2} = m_n = 1$, let $s_1 = m_1$, $s_i = s_{i-1} + m_i$, $2 \leqq i \leqq n$, and let $z_2(m)$ be the solution at $s_1, \cdots, s_{n-2}, s_{n-1} - 1$, s_{n-1} of the $(m_1, 1, \cdots, 1, m_{n-1} - 1, 1)$ conjugate boundary value problem for (1) satisfying

$$z_2(s_i) = y_i, \qquad 1 \leqq i \leqq n-2,$$
$$z_2(s_{n-1} - 1) = 0,$$
$$z_2(s_{n-1}) = y_{n-1}.$$

This time, define $S_2 = \{r \in \mathbf{R} \mid$ there is a solution $y(m)$ of (1) satisfying $y(s_i) = z_2(s_i)$, $1 \leqq i \leqq n-1$, and $y(s_n) = r\}$. Again, since $z_2(s_n) \in S_2$, S_2 is nonempty. Also, Theorem 2 implies that S_2 is an open subset of $\mathbf{R}$.

We now claim that S_2 is also closed. Assuming that S_2 is not closed, it follows that there exist $r_0 \in \bar{S}_2 \backslash S_2$ and a strictly monotone sequence $\{r_k\} \subset S_2$ such that $\lim_k r_k = r_0$. We may assume again that $r_k \uparrow r_0$ and as before, let $y_k(m)$ denote the corresponding solution of (1) satisfying

$$y_k(s_i) = z_2(s_i), \qquad 1 \leqq i \leqq n-1,$$
$$y_k(s_n) = r_k.$$

It follows from (B) that $y_k(m) > y_{k+1}(m)$ on (s_{n-2}, s_{n-1}) and $y_k(m) < y_{k+1}(m)$ on $[s_n, +\infty)$, for all $k \in \mathbf{N}$. Since $r_0 \notin S_2$ and since there exist unique solutions of $(m_1, 1, \cdots, 1, m_{n-1} - 1, 1)$ problems, Theorem 3 implies that $y_k(s_{n-1} - 1) \downarrow -\infty$, as $k \to +\infty$, and Theorem 1 implies that there exists $m_0 \in (s_n, s_n + n - 1]$ such that $y_k(m_0) \uparrow +\infty$, as $k \to +\infty$.

Let $u(m)$ be the solution at the points $s_2, \cdots, s_{n-2}, s_{n-2}+1, s_{n-1}, s_n$ of the $(m_1+1, 1, \cdots, 1, 1, m_{n-1}-1, 1)$ conjugate problem for (1) satisfying

$$u(s_i) = z_2(s_i), \qquad 2 \leqq i \leqq n-2,$$
$$u(s_{n-2}+1) = 0,$$
$$u(s_{n-1}) = z_2(s_{n-1}),$$
$$u(s_n) = r_0.$$

Since $y_k(s_{n-1}-1)\downarrow -\infty$, whereas $u(s_{n-1}) - y_k(s_{n-1}) = 0$ and $u(s_n) - y_k(s_n) > 0$ for all $k \in \mathbf{N}$, it follows that for all k sufficiently large, $u(m) - y_k(m)$ has a generalized zero at s_n. Since $y_k(m_0)\uparrow +\infty$, there exists $K \in \mathbf{N}$ such that $u(m) - y_K(m)$ has a generalized zero at s_n and a generalized zero (or zero) at some $n_0 \in (s_n, m_0]$. We also have that $u(s_i) - y_K(s_i) = 0$, $2 \leqq i \leqq n-1$, and (B) implies that $u(m) = y_K(m)$ on $[a, +\infty)$; again, this is a contradiction.

Thus, S_2 is closed and $S_2 = R$. Choosing $y_n \in S_2$, it follows that there exists a solution $y(m)$ of (1) satisfying

$$y(s_i) = y_i, \qquad 1 \leqq i \leqq n.$$

In summary, given $m_1 \in [a, +\infty)$, $m_2 = \cdots = m_{n-2} = m_n = 1$, each $(m_1, 1, \cdots, 1, m_{n-1}, 1)$ conjugate boundary value problem for (1) has a unique solution on $[a, +\infty)$.

Still assuming the induction hypotheses associated with $m_{n-1} > 1$, we assume in addition that $m_n > 1$ and that given $m_1 \in [a, +\infty)$ and $m_2 = \cdots = m_{n-2} = 1$, there exists a unique solution of each $(m_1, 1, \cdots, 1, m_{n-1}, h)$ conjugate boundary value problem, where $1 \leqq h < m_n$, for (1) on $[a, +\infty)$.

With this latter assumption, let $m_1 \in [a, +\infty)$ be given, let $m_2 = \cdots = m_{n-2} = 1$, let $s_1, \cdots, s_n$ be defined in the usual way, and let $z_3(m)$ be the solution at the points $s_1, \cdots, s_{n-2}, s_{n-1}, s_n - 1$ of the $(m_1, 1, \cdots, 1, m_{n-1}, m_n - 1)$ conjugate problem for (1) satisfying

$$z_3(s_i) = y_i, \qquad 1 \leqq i \leqq n-1,$$
$$z_3(s_n - 1) = 0.$$

Define $S_3 = \{r \in \mathbf{R} \mid$ there is a solution $y(m)$ of (1) satisfying $y(s_i) = z_3(s_i)$, $1 \leqq i \leqq n-1$, and $y(s_n) = r\}$. As before S_3 is a nonempty open subset of $\mathbf{R}$, and we claim that S_3 is also closed. Assuming again that the claim is false, let $r_0 \in \bar{S}_3 \backslash S_3$ and $\{r_k\} \subset S_3$, with $r_k \uparrow r_0$, be as in the previous considerations, and let $y_k(m)$ denote the solution of (1) satisfying

$$y_k(s_i) = z_3(s_i), \qquad 1 \leqq i \leqq n-1,$$
$$y_k(s_n) = r_k.$$

Condition (B) implies that $y_k(m) < y_{k+1}(m)$ on $(s_{n-1}, +\infty)$, for all $k \in \mathbf{N}$, and because of the existence of unique solutions of $(m_1, 1, \cdots, 1, m_{n-1}, m_n - 1)$ problems for (1) along with $r_0 \notin S_3$, Theorem 3 implies $y_k(s_n - 1)\uparrow +\infty$, as $k \to +\infty$, and Theorem 1 implies that for some $m_0 \in (s_n, s_n + n - 1]$, $y_k(m_0)\uparrow +\infty$, as $k \to +\infty$.

Now, let $u(m)$ be the solution at $s_2, \cdots, s_{n-2}, s_{n-2}+1, s_{n-1}, s_n$ of the $(m_1+1, 1, \cdots, 1, 1, m_{n-1}-1, m_n)$ boundary value problem for (1) satisfying

$$u(s_i) = z_3(s_i), \qquad 2 \leqq i \leqq n-2,$$
$$u(s_{n-2}+1) = 0,$$
$$u(s_{n-1}) = z_3(s_{n-1}),$$
$$u(s_n) = r_0.$$

Such a solution $u(m)$ exists by the primary induction hypotheses on m_{n-1} in this section of the proof. Because of the unbounded conditions on $\{y_k(s_n-1)\}$ and $\{y_k(m_0)\}$, while $u(s_n)>y_k(s_n)$, for all $k\in\mathbf{N}$, there exists $K\in\mathbf{N}$ such that $u(m)-y_K(m)$ has a generalized zero at s_n and a generalized zero at some $n_0\in(s_n, m_0]$. Moreover, $u(s_i)-y_K(s_i)=0$, $2\leqq i\leqq n-1$, from which it follows that $u(m)=y_K(m)$ on $[a,+\infty)$, a contradiction.

Consequently, S_3 is closed, $S_3=\mathbf{R}$, and choosing $y_n\in S_3$, the corresponding solution $y(m)$ of (1) satisfying $y(s_n)=y_n$ is the desired solution. In particular, given $m_1\in[a,+\infty)$, $m_2=\cdots=m_{n-2}=1$, and $m_n\geqq 1$, each $(m_1,1,\cdots,1,m_{n-1},m_n)$ conjugate boundary value problem for (1) has a unique solution on $[a,+\infty)$.

This completes the induction on m_{n-1}. That is, given $m_1\in[a,+\infty)$, $m_2=\cdots=m_{n-2}=1$, and m_{n-1}, $m_n\geqq 1$, each $(m_1,1,\cdots,1,m_{n-1},m_n)$ conjugate boundary value problem for (1) has a unique solution on $[a,+\infty)$.

To illustrate the pattern of the induction better, we will give some of the details of the four steps involved in the induction on m_{n-2}.

For this section of the proof, our outstanding assumption is that $m_{n-2}>1$, and that given $m_1\in[a,+\infty)$, $m_2=\cdots=m_{n-3}=1$, and m_{n-1}, $m_n\geqq 1$, there exists a unique solution of each $(m_1,1,\cdots,1,k,m_{n-1},m_n)$ conjugate boundary value problem, where $1\leqq k<m_{n-2}$, for (1) on $[a,+\infty)$. Under that assumption, we will be concerned with solutions of $(m_1,1,\cdots,1,m_{n-2},1,1)$ followed by $(m_1,1,\cdots,1,m_{n-2},1,m_n)$, $m_n>1$, followed by $(m_1,1,\cdots,1,m_{n-2},m_{n-1},1)$, $m_{n-1}>1$, followed by $(m_1,1,\cdots,1,m_{n-2},m_{n-1},m_n)$, m_{n-1}, $m_n>1$, boundary value problems for (1).

Let $m_1\in[a,+\infty)$, let $m_2=\cdots=m_{n-3}=m_{n-1}=m_n=1$, let s_i, $1\leqq i\leqq n$, be as usual, and let $z_4(m)$ be the solution of the $(m_1,1,\cdots,1,m_{n-2}-1,1,1)$ conjugate boundary value problem for (1) satisfying

$$z_4(s_i)=y_i,\qquad 1\leqq i\leqq n-3,$$
$$z_4(s_{n-2}-1)=0,$$
$$z_4(s_i)=y_i,\qquad i=n-2,n-1.$$

Defining $S_4=\{r\in\mathbf{R}\,|\,$ there is a solution $y(m)$ of (1) satisfying $y(s_i)=z_4(s_i)$, $1\leqq i\leqq n-1$, and $y(s_n)=r\}$, S_4 is nonempty and open.

If we assume S_4 is not closed, then let r_0 and $\{r_k\}$, with $r_k\uparrow r_0$, be as usual and let $y_k(m)$ denote the corresponding solution of (1). It follows in this case that $y_k(s_{n-2}-1)\uparrow+\infty$, as $k\to+\infty$, and for some $m_0\in(s_n, s_n+n-1]$, $y_k(m_0)\uparrow+\infty$, as $k\to+\infty$.

Denoting by $u(m)$ the solution of the $(m_1+1,1,\cdots,1,1,m_{n-2}-1,1,1)$ problem for (1) satisfying

$$u(s_i)=y_i,\qquad 2\leqq i\leqq n-3,$$
$$u(s_{n-3}+1)=0,$$
$$u(s_i)=y_i,\qquad i=n-2,n-1,$$
$$u(s_n)=r_0,$$

it follows that for some $K\in\mathbf{N}$, $u(m)-y_K(m)$ has a generalized zero at s_n, a generalized zero at some $n_0\in(s_n, m_0]$, and zeros at s_i, $2\leqq i\leqq n-1$. Again, we contradict (B), and hence S_4 is closed. Select $y_n\in S_4$ and the corresponding solution is the desired solution of the $(m_1,1,\cdots,1,m_{n-2},1,1)$ problem for (1).

In addition to our assumptions on $m_{n-2}>1$, we assume that $m_n>1$ and that given $m_1\in[a,+\infty)$, $m_2=\cdots=m_{n-3}=m_{n-1}=1$, each $(m_1,1,\cdots,1,m_{n-2},1,h)$ conjugate boundary value problem, where $1\leqq h<m_n$, for (1) has a unique solution on $[a,+\infty)$.

Given $m_1 \in [a, +\infty)$ and $m_2 = \cdots = m_{n-3} = m_{n-1} = 1$ with s_i, $1 \leqq i \leqq n$, as usual, let $z_5(m)$ be the solution of the $(m_1, 1, \cdots, 1, m_{n-2}, 1, m_n - 1)$ problem for (1) satisfying

$$z_5(s_i) = y_i, \qquad 1 \leqq i \leqq n-1,$$
$$z_5(s_n - 1) = 0.$$

Defining S_5 in the standard way, S_5 is nonempty and open.

If we assume S_5 is not closed, then let r_0 and $\{r_k\}$, with $r_k \uparrow r_0$, be as usual, and let $y_k(m)$ be the appropriate solution of (1). By the existence of unique solutions of $(m_1, 1, \cdots, 1, m_{n-2}, 1, m_n - 1)$ problems for (1), we have that $y_k(s_n - 1) \uparrow +\infty$. Also, $y_k(m_0) \uparrow +\infty$, where m_0 is as usual. In this case, now let $u(m)$ be the solution of the $(m_1 + 1, 1, \cdots, 1, m_{n-2} - 1, 1, m_n)$ problem for (1) satisfying

$$u(s_i) = z_5(s_i), \qquad 2 \leqq i \leqq n-3,$$
$$u(s_{n-3} + 1) = 0,$$
$$u(s_i) = z_5(s_i), \qquad i = n-2, n-1,$$
$$u(s_n) = r_0.$$

In this case, there exists $K \in \mathbf{N}$ such that $u(m) - y_K(m)$ has a generalized zero at s_n, a generalized zero at some $n_0 \in (s_n, m_0]$, and zeros at s_i, $2 \leqq i \leqq n-1$, the usual contradiction.

Thus S_5 is closed, and we conclude the existence of unique solutions of $(m_1, 1, \cdots, 1, m_{n-2}, 1, m_n)$ conjugate boundary value problems for (1) on $[a, +\infty)$.

In addition to the primary induction hypotheses on m_{n-2}, we assume now that $m_{n-1} > 1$ and that given $m_1 \in [a, +\infty)$, $m_2 = \cdots = m_{n-3} = 1$, and $m_n \geqq 1$, there exists a unique solution of each $(m_1, 1, \cdots, 1, m_{n-2}, l, m_n)$ conjugate boundary value problem, where $1 \leqq l < m_{n-1}$, for (1) on $[a, +\infty)$.

In this case, we let $m_1 \in [a, +\infty)$, $m_2 = \cdots = m_{n-3} = m_n = 1$, and we let $z_6(m)$ be the solution of the $(m_1, 1, \cdots, 1, m_{n-2}, m_{n-1} - 1, 1)$ problem for (1) satisfying

$$z_6(s_i) = y_i, \qquad 1 \leqq i \leqq n-2,$$
$$z_6(s_{n-1} - 1) = 0,$$
$$z_6(s_{n-1}) = y_{n-1}.$$

The corresponding set S_6 will be nonempty and open. Repeating the pattern, we assume S_6 is not closed and make the usual arguments using r_0, $\{r_k\}$, and the corresponding solutions of (1). In this case $y_k(s_{n-1} - 1) \downarrow -\infty$, as $k \to +\infty$, and for some $m_0 \in (s_n, s_n + n - 1]$, $y_k(m_0) \uparrow +\infty$, as $k \to +\infty$.

With $u(m)$ the solution at $s_2, \cdots, s_{n-3}, s_{n-3} + 1, s_{n-2}, s_{n-1}, s_n$ of the $(m_1 + 1, 1, \cdots, 1, m_{n-2} - 1, m_{n-1}, 1)$ boundary value problem for (1) satisfying

$$u(s_i) = z_6(s_i), \qquad 2 \leqq i \leqq n-3,$$
$$u(s_{n-3} + 1) = 0,$$
$$u(s_i) = z_6(s_i), \qquad i = n-2, n-1,$$
$$u(s_n) = r_0,$$

it follows that for some $K \in \mathbf{N}$, $u(m) - y_K(m)$ has a generalized zero at s_n, a generalized zero at some $n_0 \in (s_n, m_0]$, and zeros at s_i, $2 \leqq i \leqq n-1$. This is a contradiction to (B), and hence S_6 is closed; consequently each $(m_1, 1, \cdots, 1, m_{n-2}, m_{n-1}, 1)$ problem for (1) has a unique solution.

For the final step under the primary induction hypotheses on $m_{n-2}>1$ and the induction hypotheses on $m_{n-1}>1$, assume in addition that $m_n>1$ and that given $m_1\in[a,+\infty)$ and $m_2=\cdots=m_{n-3}=1$, there exists a unique solution of each $(m_1,1,\cdots,1,m_{n-2},m_{n-1},h)$ conjugate boundary value problem, where $1\leqq h<m_n$, for (1) on [a, $+\infty$).

To complete the argument, let $m_1\in[a,+\infty)$ be given, let $m_2=\cdots=m_{n-3}=1$, and let $z_7(m)$ be the solution of the $(m_1,1,\cdots,1,m_{n-2},m_{n-1},m_n-1)$ boundary value problem for (1) satisfying

$$z_7(s_i)=y_i,\qquad 1\leqq i\leqq n-1$$
$$z_7(s_n-1)=0.$$

Defining the nonempty open set S_7 in our standard manner, and making the usual assumption that S_7 is not closed, let r_0, $\{r_k\}$, and $y_k(m)$ be the appropriate values and solutions. We can argue that $y_k(s_n-1)\uparrow+\infty$ and $y_k(m_0)\uparrow+\infty$, for some $m_0\in(s_n,s_n+n-1]$. If $u(m)$ is the solution of the $(m_1+1,1,\cdots,1,m_{n-2}-1,m_{n-1},m_n)$ conjugate boundary value problem for (1) satisfying

$$u(s_i)=z_7(s_i),\qquad 2\leqq i\leqq n-3,$$
$$u(s_{n-3}+1)=0,$$
$$u(s_i)=z_7(s_i),\qquad i=n-2,n-1,$$
$$u(s_n)=r_0,$$

then there exists $K\in\mathbf{N}$ such that $u(m)-y_K(m)$ has a generalized zero at s_n, a generalized zero at some $n_0\in(s_n,m_0]$, and zeros at s_i, $2\leqq i\leqq n-1$. This contradicts (B); hence S_7 is closed, and as in each of the above cases, the $(m_1,1,\cdots,1,m_{n-2},m_{n-1},m_n)$ problem has a unique solution.

That concludes the induction on m_{n-2}; in particular, given $m_1\in[a,+\infty)$, $m_2=\cdots=m_{n-3}=1$, and $m_{n-2},m_{n-1},m_n\geqq 1$, each $(m_1,1,\cdots,1,m_{n-2},m_{n-1},m_n)$ conjugate boundary value problem for (1) has a unique solution on $[a,+\infty)$.

Although our above arguments exhibit the entire pattern for the induction scheme in obtaining solutions of the boundary value problems, for completeness we will include some of the details involved in the general induction step. To that end, assume that $2\leqq p\leqq n-3$, that $m_p>1$, and that given $m_1\in[a,+\infty)$, $m_2=\cdots=m_{p-1}=1$, and $m_{p+1},\cdots,m_n\geqq 1$, there exists a unique solution of each $(m_1,1,\cdots,1,k,m_{p+1},\cdots,m_n)$ conjugate boundary value problem, where $1\leqq k<m_p$, for (1) on $[a,+\infty)$.

Under that assumption we will be concerned with establishing the existence of solutions of $(m_1,1,\cdots,1,m_p,m_{p+1},\cdots,m_n)$ problems by proceeding through 2^{n-p} inductive steps, wherein we induct on $m_n,m_{n-1},\cdots,m_{p+1}$, following the pattern in the above parts of this proof. Now for each one of these 2^{n-p} steps, there exist natural numbers $1\leqq j_1<j_2<\cdots<j_s\leqq n-p$, such that we are concerned with the $(2^{j_s}+2^{j_{s-1}}+\cdots+2^{j_2}+2^{j_1})$st inductive step or with the $(2^{j_s}+2^{j_{s-1}}+\cdots+2^{j_2}+2^{j_1}+1)$st inductive step.

(a) In the case of the $(2^{j_s}+2^{j_{s-1}}+\cdots+2^{j_2}+2^{j_1})$st inductive step, our concern is with showing the existence of solutions of $(m_1,1,\cdots,1,m_p,1,\cdots,1,m_{n-j_s},1,\cdots,1,m_{n-j_{s-1}},1,\cdots,1,m_{n-j_2},1,\cdots,1,m_{n-j_1+1},m_{n-j_1+2},\cdots,m_n)$ problems for (1), where the reader realizes the significance of the entries in the n-tuple from previous arguments.

For this problem, in addition to appropriate assumptions from preceding steps on $m_p,m_{n-j_s},\cdots,m_{n-j_2},m_{n-j_1+1},\cdots,m_{n-1}$, our assumptions are that $m_n>1$ and that given $m_1\in[a,+\infty)$, each $(m_1,1,\cdots,1,m_p,1,\cdots,1,m_{n-j_s},1,\cdots,1,m_{n-j_{s-1}},1,\cdots,1,$

$m_{n-j_2}, 1, \cdots, 1, m_{n-j_1+1}, \cdots, m_{n-1}, h)$ conjugate boundary value problem, where $1 \leqq h < m_n$, for (1) has a unique solution on $[a, +\infty)$.

Given $m_1 \in [a, +\infty)$, let $z(m)$ be the solution at the points $s_1, \cdots, s_{n-1}, s_n - 1$ of the $(m_1, 1, \cdots, 1, m_p, 1, \cdots, 1, m_{n-j_s}, 1, \cdots, 1, m_{n-j_{s-1}}, 1, \cdots, 1, m_{n-j_2}, 1, \cdots, 1, m_{n-j_1+1}, \cdots, m_{n-1}, m_n - 1)$ conjugate boundary value problem for (1) satisfying

$$z(s_i) = y_i, \qquad 1 \leqq i \leqq n-1,$$

$$z(s_n - 1) = 0.$$

The set $S = \{r \in \mathbf{R} \mid$ there is a solution $y(m)$ of (1) satisfying $y(s_i) = z(s_i)$, $1 \leqq i \leqq n-1$, and $y(s_n) = r\}$ is nonempty and open.

If we assume as above that S is not closed, and if r_0, $\{r_k\}$, with $r_k \uparrow r_0$, and $y_k(m)$ are also as in our previous arguments, then it follows from (B) that $y_k(m) < y_{k+1}(m)$, for each $k \in \mathbf{N}$, on $(s_{n-1}, +\infty)$. In complete analogy also, $y_k(s_n - 1) \uparrow +\infty$, as $k \to +\infty$, and for some $m_0 \in (s_n, s_n + n - 1]$, $y_k(m_0) \uparrow +\infty$, as $k \to +\infty$.

If $u(m)$ is the solution at the points $s_2, s_3, \cdots, s_{p-1}, s_{p-1}+1, s_p, s_{p+1}, \cdots, s_{n-1}, s_n$ of the $(m_1+1, 1, \cdots, 1, 1, m_p - 1, 1, \cdots, 1, m_{n-j_s}, 1, \cdots, 1, m_{n-j_{s-1}}, 1, \cdots, 1, m_{n-j_2}, 1, \cdots, 1, m_{n-j_1+1}, \cdots, m_{n-1}, m_n)$ conjugate boundary value problem for (1) satisfying

$$u(s_i) = z(s_i), \qquad 2 \leqq i \leqq p-1,$$

$$u(s_{p-1} + 1) = 0,$$

$$u(s_i) = z(s_i), \qquad p \leqq i \leqq n-1,$$

$$u(s_n) = r_0,$$

then we can argue that, for some $K \in \mathbf{N}$, $u(m) - y_K(m)$ has a generalized zero at s_n, a generalized zero at some $n_0 \in (s_n, m_0]$, and zeros at s_i, $2 \leqq i \leqq n-1$. This contradicts (B), and it follows that S is also closed. For $y_n \in S$, the corresponding solution $y(m)$ is the desired solution. This completes this case.

(b) For the case of the $(2^{j_s} + 2^{j_{s-1}} + \cdots + 2^{j_2} + 2^{j_1} + 1)$st inductive step, we are concerned with the existence of solutions of $(m_1, 1, \cdots, 1, m_p, 1, \cdots, 1, m_{n-j_s}, 1, \cdots, 1, m_{n-j_{s-1}}, 1, \cdots, 1, m_{n-j_2}, 1, \cdots, 1, m_{n-j_1}, 1, \cdots, 1)$ problems for (1).

As in (a), in addition to appropriate assumptions on $m_p, m_{n-j_s}, \cdots, m_{n-j_2}$, our assumptions are that $m_{n-j_1} > 1$ and that given $m_1 \in [a, +\infty)$, each $(m_1, 1, \cdots, 1, m_p, 1, \cdots, 1, m_{n-j_s}, 1, \cdots, 1, m_{n-j_{s-1}}, 1, \cdots, 1, m_{n-j_2}, 1, \cdots, 1, l, m_{n-j_1+1}, m_{n-j_1+2}, \cdots, m_n)$ conjugate boundary value problem, where $1 \leqq l < m_{n-j_1}$, for (1) has a unique solution on $[a, +\infty)$.

Given $m_1 \in [a, +\infty)$, let $z(m)$ be the solution at the points $s_1, \cdots, s_{n-j_1-1}, s_{n-j_1} - 1, s_{n-j_1}, s_{n-j_1+1}, \cdots, s_{n-2}, s_{n-1}$ of the $(m_1, 1, \cdots, 1, m_p, 1, \cdots, 1, m_{n-j_s}, 1, \cdots, 1, m_{n-j_{s-1}}, 1, \cdots, 1, m_{n-j_2}, 1, \cdots, 1, m_{n-j_1} - 1, 1, \cdots, 1)$ conjugate boundary value problem for (1) satisfying

$$z(s_i) = y_i, \qquad 1 \leqq i \leqq n - j_1 - 1,$$

$$z(s_{n-j_1} - 1) = 0,$$

$$z(s_i) = y_i, \qquad n - j_1 \leqq i \leqq n-1.$$

If $S = \{r \in \mathbf{R} \mid$ there is a solution $y(m)$ of (1) satisfying $y(s_i) = z(s_i)$, $1 \leqq i \leqq n-1$, and $y(s_n) = r\}$, then S is a nonempty open subset of $\mathbf{R}$.

If we assume S is not closed, and if $r_0, \{r_k\}$, with $r_k \uparrow r_0$, and $y_k(m)$ are the standard points and solutions, then $y_k(m) < y_{k+1}(m)$ on $[s_n, +\infty)$, for all $k \in \mathbf{N}$. Moreover, since $y_k(s_i) = z(s_i)$, $n - j_1 \leqq i \leqq n-1$, it follows from (B) that either

(i) $y_k(s_{n-j_1}-1) < y_{k+1}(s_{n-j_1}-1)$, for each $k \in \mathbf{N}$, if j_1 is even, or

(ii) $y_k(s_{n-j_1}-1) > y_{k+1}(s_{n-j_1}-1)$, for each $k \in \mathbf{N}$, if j_1 is odd.

Since the arguments for both cases are analogous, we may assume case (i) is the situation. In this case, since $r_0 \notin S$, it follows from Theorem 3 that $y_k(s_{n-j_1}-1) \uparrow +\infty$, as $k \to +\infty$. Furthermore, it is again the case that, for some $m_0 \in (s_n, s_n + n - 1]$, $y_k(m_0) \uparrow +\infty$, as $k \to +\infty$.

Now let $u(m)$ be the solution at the points $s_2, \cdots, s_{p-1}, s_{p-1}+1, s_p, s_{p+1}, \cdots, s_{n-1}, s_n$ of the $(m_1+1, 1, \cdots, 1, 1, m_p - 1, 1, \cdots, 1, m_{n-j_s}, 1, \cdots, 1, m_{n-j_{s-1}}, 1, \cdots, 1, m_{n-j_2}, 1, \cdots, 1, m_{n-j_1}, 1, \cdots, 1)$ conjugate boundary value problem for (1) satisfying

$$
\begin{aligned}
&u(s_i) = z(s_i), \qquad 2 \leqq i \leqq p-1,\\
&u(s_{p-1}+1) = 0,\\
&u(s_i) = z(s_i), \qquad p \leqq i \leqq n-1,\\
&u(s_n) = r_0.
\end{aligned}
$$

As with previous similar cases, there exists $K \in \mathbf{N}$ such that $u(m) - y_K(m)$ has a generalized zero at s_n, a generalized zero at some $n_0 \in (s_n, m_0]$, and zeros at s_i, $2 \leqq i \leqq n-1$. This is a contradiction to (B), and we conclude that S is closed; hence $S = \mathbf{R}$. Choosing $y_n \in S$, the associated solution $y(m)$ is the unique solution of (1) that we sought. This completes case (b).

The proof is complete. □

Remark. We remark here that the half-line $[a, +\infty)$ is not necessary. In particular, the results can be extended to a finite interval $[a, b+n]$, where b is the right-most point at which conditions are specified, so that our application of Theorem 1 can still be made in the arguments.

REFERENCES

[1] R. P. AGARWAL, *On multipoint boundary value problems for discrete equations*, J. Math. Anal. Appl., 96 (1983), pp. 520-534.

[2] ———, *Initial-value methods for discrete boundary value problems*, J. Math. Anal. Appl., 100 (1984), pp. 513-529.

[3] ———, *Difference calculus with applications to difference equations*, in Proc. Conference on General Inequalities 4, Oberwolfach, Internat. Ser. Numer. Math., 71 (1984), pp. 95-110.

[4] ———, *Initial and boundary value problems for nth order difference equations*, Math. Slovaca, 36 (1986), pp. 39-47.

[5] R. P. AGARWAL AND R. C. GUPTA, *A new shooting method for multi-point boundary value problems*, J. Math. Anal. Appl., 112 (1985), pp. 210-220.

[6] C. AHLBRANDT AND J. HOOKER, *A variational view of nonoscillation theory for linear difference equations*, in Part 2 of Proc. Twelfth and Thirteenth Midwest Conferences on Differential and Integral Equations, J. Henderson, ed., University of Missouri, Rolla, MO, 1985, pp. 1-21.

[7] ———, *Riccati transformations and principal solutions of discrete linear systems*, in Proc. 1984 Workshop on Spectral Theory of Sturm-Liouville Differential Operators, ANL-84-87, Argonne National Laboratory, Argonne, IL, 1984.

[8] ———, *Disconjugacy criteria for second order linear difference equations*, in Proc. 1984 Edmonton Conference on Qualitative Properties of Differential Equations, W. Allegretto and G. J. Butler, eds., University of Alberta, Edmonton, Alberta, Canada, 1986, pp. 15-26

[9] P. W. ELOE, *Difference equations and multipoint boundary value problems*, Proc. Amer. Math. Soc., 86 (1982), pp. 253-259.

[10] P. W. ELOE, *A boundary value problem for a system of difference equations*, Nonlinear Anal., 7 (1983), pp. 813–820.
[11] ———, *Criteria for right disfocality of linear difference equations*, J. Math. Anal. Appl., 120 (1986), pp. 610–621.
[12] ———, *A comparison theorem for linear difference equations*, Proc. Amer. Math. Soc., to appear.
[13] ———, *Eventual disconjugacy and right disfocality of linear difference equations*, Bull. Canad. Math., to appear.
[14] D. HANKERSON, *Boundary value problems for n-th order difference equations*, Ph.D. dissertation, University of Nebraska, Lincoln, NE, 1987.
[15] D. HANKERSON AND A. PETERSON, *On a theorem of Elias for difference equations*, in Proc. VII International Conference on Nonlinear Analysis and Application, V. Lakshmikantham, ed., Marcel Dekker, New York, Basel, 1987.
[16] ———, *A classification of the solutions of a difference equation according to their behavior at infinity*, J. Math. Anal. Appl., to appear.
[17] P. HARTMAN, *On n-parameter families and interpolation problems for nonlinear ordinary differential equations*, Trans. Amer. Math. Soc., 154 (1971), pp. 201–226.
[18] ———, *Difference equations: Disconjugacy, principal solutions, Green's functions, complete monotonicity*, Trans. Amer. Math. Soc., 246 (1978), pp. 1–30.
[19] J. HOOKER, M. K. KWONG, AND W. PATULA, *Riccati type transformations for second order linear difference equations* II, J. Math. Anal. Appl., 107 (1985), pp. 182–196.
[20] J. HOOKER AND W. PATULA, *Riccati type transformations for second-order linear difference equations*, J. Math. Anal. Appl., 82 (1981), pp. 451–462.
[21] ———, *A second-order nonlinear difference equation: oscillation and asymptotic behavior*, J. Math. Anal. Appl., 91 (1983), pp. 9–29.
[22] ———, *Growth and oscillation properties of solutions of fourth order linear difference equations*, J. Australian Math. Soc., Ser. B, 26 (1985), pp. 310–328.
[23] G. KLAASEN, *Existence theorems for boundary value problems for nth order ordinary differential equations*, Rocky Mountain J. Math., 3 (1973), pp. 457–472.
[24] W. PATULA, *Growth, oscillation and comparison theorems for second-order linear difference equations*, SIAM J. Math. Anal., 10 (1979), pp. 1272–1279.
[25] A. PETERSON, *Boundary value problems for an nth order linear difference equation*, SIAM J. Math. Anal., 15 (1984), pp. 124–132.
[26] ———, *On $(k, n-k)$-disconjugacy for linear difference equations*, in Proc. 1984 Edmonton Conference on Qualitative Properties of Differential Equations, W. Allegretto and G. J. Butler, eds., University of Alberta, Edmonton, Alberta, Canada, 1986, pp. 329–337.
[27] ———, *Boundary value problems and Green's function for linear difference equations*, in Part 1 of Proc. Twelfth and Thirteenth Midwest Conferences on Differential and Integral Equations, J. Henderson, ed., University of Missouri, Rolla, MO, 1985, pp. 79–100.
[28] ———, *Green's functions for $(k, n-k)$-boundary value problems for linear difference equations*, J. Math. Anal. Appl., 124 (1987), pp. 127–138.
[29] ———, *Existence and uniqueness theorems for nonlinear difference equations*, J. Math. Anal. Appl., 125 (1987), pp. 185–191.
[30] J. RODRIGUEZ, *On nonlinear discrete boundary problems*, J. Math. Anal. Appl., 114 (1986), pp. 398–408.
[31] B. SMITH AND W. TAYLOR, JR., *Oscillatory and asymptotic behavior of certain fourth order difference equations*, Rocky Mountain J. Math., 16 (1986), pp. 403–406.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 479-493, March 1989

015

UNIFORM APPROXIMATION OF SINGULAR PERTURBATION PROBLEMS HAVING SINGULAR REGULAR EXPANSIONS*

WALTER G. KELLEY†

Abstract. The singular perturbation problem $\varepsilon^2 u'' = h(x, u, \varepsilon)$, $u(a) = u(b) = 0$ is studied under the assumption that h vanishes for some $x \in [a, b]$. Uniform asymptotic approximations are obtained for solutions exhibiting boundary layer behavior. Approximate Green's functions are constructed to show existence of solutions and verify formal approximations.

Key words. singular perturbations, semilinear boundary value problem, singular expansion, boundary layer, asymptotic approximations

AMS(MOS) subject classifications. 34E15, 34E05

1. Introduction.

Elliptic singularly perturbed boundary value problems of the form

$$\varepsilon^2 \nabla^2 u = f(x, u, \varepsilon), \qquad x \in \Omega, \tag{1.1}$$

$$u(x) = 0, \qquad x \in \partial\Omega, \tag{1.2}$$

where ε is a small positive parameter, have been studied by Fife [2], van Harten [3] and others under the following hypotheses:

(a) There is a smooth u_0 so that $f(x, u_0(x), 0) = 0$ for $x \in \Omega \cup \partial\Omega$.

(b) $f_u(x, u_0(x), 0) > 0$ for $x \in \Omega \cup \partial\Omega$.

(c) $\int_{u_0(x)}^{A} f(x, s, 0)\, ds > 0$ for $x \in \partial\Omega$ and A between $u_0(x)$ and 0.

They have shown that, for sufficiently small values of ε, problem (1.1), (1.2) has a solution u_ε with a boundary layer of width ε along $\partial\Omega$ and that asymptotic approximations of all orders can be constructed.

Recently, we have obtained equations similar to (1.1) while using regularization methods for optimization problems that arise in the identification of parameters. However, in these problems the function f vanishes identically for some x in the domain so that hypothesis (b) is violated. Note that (c) is also violated if x happens to be a boundary point.

In this paper, as a first step in analyzing such problems, we consider ordinary differential equations

$$\varepsilon^2 u'' = h(x, u, \varepsilon) \tag{1.3}$$

with Dirichlet boundary conditions, where $h(0, u, 0) = 0$ for all u and $x = 0$ belongs to the domain of interest, and show that there is a solution with boundary layers at both endpoints. The exact behavior of the solution will be obtained by constructing and verifying a uniform asymptotic approximation. The analysis has some connection with classical turning point theory for linear equations (see Wasow [10]). However, our hypotheses rule out the possibility that solutions exhibit rapid oscillations. Cochran [1] has given a method for constructing solutions for linear boundary value problems under similar hypotheses.

In § 3 we consider first the case when $x = 0$ is in the interior of the interval. One interesting aspect of the computation is that the regular expansion has a singularity at $x = 0$. Nevertheless, a composite approximation is easily constructed and is verified

* Received by the editors August 24, 1987; accepted for publication May 2, 1988.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019.

by the use of an approximate Green's function. We do not consider higher-order approximations.

In the final section, we treat the case where $x=0$ is a boundary point. Despite the fact that h vanishes at the boundary, it is shown that an integral condition similar to hypothesis (c) suffices to obtain a solution with boundary layer behavior. We also verify that a lowest-order approximation consisting of the reduced solution with boundary layer corrections is uniformly valid. In contrast to the situation where hypotheses (a)-(c) hold, the boundary layer at $x=0$ is found to be thicker than the other one.

2. Preliminaries. In the following sections, we will use the method of approximate Green's functions introduced in van Harten and Vader-Burger [4] to show the existence of solutions and to verify formal approximations. Consequently, this section will summarize the results of [4] as they apply to (1.3) with Dirichlet boundary conditions

$$u(a)=u(b)=0. \tag{2.1}$$

We will assume throughout that all functions are smooth enough to make the arguments valid, but class C^2 is sufficient in all cases.

Suppose a formal approximation $A(x,\varepsilon)$ has been obtained, in the sense that

$$\varepsilon^2 A''-h(x,A,\varepsilon)=E(x,\varepsilon)=\mathcal{O}(\varepsilon^{\alpha}) \tag{2.2}$$

for some $\alpha>0$ and all $x\in[a,b]$ and A satisfies (2.1). Then

$$L_\varepsilon(v)=\varepsilon^2 v''-h_u(x,A,\varepsilon)v \tag{2.3}$$

is the linearization of (2.2) about A and the residual nonlinear operator is given by

$$N_\varepsilon(v)=-\int_0^1 (1-s)\frac{d^2}{ds^2}h(x,A+sv,\varepsilon)\,ds. \tag{2.4}$$

Let $\|\ \|$ denote the supremum norm on $C[a,b]$.

DEFINITION. A function $\mathrm{Gr}(x,t,\varepsilon)$ is said to be an "approximate Green's function" for (1.3), (2.1) if it satisfies the following conditions:

(1) $\mathrm{Gr}\in C([a,b]^2)$ and Gr is smooth except across the diagonal $x=t$;

(2) The jump in Gr_x at $x=t$ satisfies

$$[\mathrm{Gr}_x]_{x=t}=1+\mathcal{O}(\varepsilon);$$

(3) There is a $\nu>0$ so that

$$\int_a^b |L_\varepsilon\,\mathrm{Gr}(x,t)-\delta(x-t)|\,dt=\mathcal{O}(\varepsilon^{\nu})$$

uniformly for $x\in[a,b]$, where δ is the Dirac delta function;

(4) Gr satisfies $\mathrm{Gr}(a,t,\varepsilon)=\mathrm{Gr}(b,t,\varepsilon)=0$ for all $t\in[a,b]$.

If Gr is an approximate Green's function, we define an approximate right inverse L_ε^{-1} for L_ε by

$$L_\varepsilon^{-1}u(x)=\int_a^b \mathrm{Gr}(x,t)u(t)\,dt. \tag{2.5}$$

The proof of the following theorem in [4] is based on a contraction mapping argument.

THEOREM 1. *Suppose that A satisfies* (2.1), (2.2), N_ε *and* L_ε^{-1} *are given by* (2.4), *and* (2.5), *respectively, where* Gr *is an approximate Green's function, and*

$$\|N_\varepsilon L_\varepsilon^{-1}u_1 - N_\varepsilon L_\varepsilon^{-1}u_2\| \leqq \frac{\rho}{\beta(\varepsilon)}\|u_1 - u_2\| \tag{2.6}$$

for $\|u_1\|, \|u_2\| < \rho$, $0 < \rho < \bar{\rho}(\varepsilon)$, ε *sufficiently small and* $\beta(\varepsilon) > 0$. *If*

$$\|E\| \leqq \min\{\bar{\rho}(\varepsilon)/4, \beta(\varepsilon)/16\}, \tag{2.7}$$

then for $\varepsilon > 0$ *sufficiently small,* (1.3), (2.1), *has a locally unique solution* $u(x, \varepsilon)$ *so that*

$$\|u(x, \varepsilon) - A(x, \varepsilon)\| \leqq 4|L_\varepsilon^{-1}|\,\|E\|,$$

where $|\ |$ *is the norm of the operator* L_ε^{-1} *as a mapping from* $C[a, b]$ *into* $C[a, b]$.

This theorem with minor modification holds as well for a variety of perturbation problems and Banach spaces.

Now the calculations in [3] imply that

$$\|N_\varepsilon v_1 - N_\varepsilon v_2\| = \mathcal{O}(\rho)\|v_1 - v_2\|$$

for ρ and ε small, where ρ is independent of ε, so (2.6) is true with $\beta = C/|L_\varepsilon^{-1}|$. Then (2.7) is satisfied if

$$\|E\| \times |L_\varepsilon^{-1}| \to 0 \quad \text{as } \varepsilon \to 0. \tag{2.8}$$

COROLLARY. *If the error term E in* (2.2) *satisfies* (2.8), *then for small* ε, (1.3), (2.1) *has a locally unique solution* $u(x, \varepsilon)$ *so that*

$$\|u - A\| \leqq 4|L_\varepsilon^{-1}|\,\|E\|. \tag{2.9}$$

3. Problems with an interior singularity. Consider now the problem

$$\varepsilon^2 u'' = x^m f(x, u) + \varepsilon^2 g(x, u, \varepsilon), \tag{3.1}$$

$$u(-1) = u(1) = 0, \tag{3.2}$$

where m is an even positive integer. More general problems could be considered; for example, we could replace x^m by $|x|^m$, where m is allowed to be odd. We will assume the following:

(a) There is a smooth u_0 so that $f(x, u_0(x)) = 0$, $-1 \leqq x \leqq 1$;

(b) $f_u(x, u_0) > k^2 > 0$, $-1 \leqq x \leqq 1$;

(c) $\int_{u_0(-1)}^{\theta} f(-1, s)\,ds > 0$ for θ between $u_0(-1)$ and 0;

(d) $\int_{u_0(1)}^{\theta} f(1, s)\,ds > 0$ for θ between $u_0(1)$ and 0.

The first step is to construct an approximation of the solution of (3.1), (3.2) valid near $x = 0$. If a regular expansion $u = u_0(x) + \varepsilon^2 u_1(x) + \cdots$ is substituted into (3.1), we find that

$$u_1(x) = \frac{u_0''(x) - g(x, u_0, 0)}{x^m f_u(x, u_0)}, \tag{3.3}$$

which is generally singular at $x = 0$. We consider only the worst case, namely that

(e) $F(x) \equiv u_0''(x) - g(x, u_0(x), 0) \neq 0$ at $x = 0$.

To correct the singularity, we try as an approximation near zero

$$u_0(x) + \varepsilon^{4/(m+2)} v(\xi, \varepsilon) \tag{3.4}$$

with $\xi = x\varepsilon^{-2/(m+2)}$. If we put (3.4) into (3.1), we find that v satisfies the non-homogeneous equation

$$\frac{d^2 v}{d\xi^2} = \xi^m f_u(0, u_0(0))v - F(0). \tag{3.5}$$

To obtain matching with the first two terms of the regular expansion, we want a solution of (3.5) so that

$$v(\xi, \varepsilon) \sim \frac{F(0)}{\xi^m f_u(0, u_0(0))} \quad \text{as } \xi \to \pm\infty. \tag{3.6}$$

Now let $\eta = \xi f_u(0, u_0(0))^{1/(m+2)}$ and write (3.5) in terms of the independent variable η:

$$\frac{d^2v}{d\eta^2} = \eta^m v - \frac{F(0)}{f_u(0, u_0(0))^{2/(m+2)}}. \tag{3.7}$$

It is well known (see, for example, [11]) that the homogeneous portion of (3.7) has the solution

$$w(\eta) = \eta^{1/2} K_{1/(m+2)}\left(\frac{2\eta^{(m+2)/2}}{m+2}\right), \tag{3.8}$$

for $\eta \geqq 0$, where $K_{1/(m+2)}$ is a modified Bessel function; furthermore, w can be continued to an analytic function defined for all η. Note that $w(-\eta)$ is also a solution of the homogeneous part of (3.7).

Define

$$v(\eta) = \frac{F(0)}{\mathrm{Wr}}\left(\frac{1}{f_u(0, u_0(0))}\right)^{2/(m+2)} \left\{ w(-\eta) \int_\eta^\infty w(t)\, dt + w(\eta) \int_{-\infty}^\eta w(-t)\, dt \right\}, \tag{3.9}$$

where Wr is the Wronskian of $w(\eta)$, $w(-\eta)$:

$$\mathrm{Wr} = -2w(0)w'(0) > 0.$$

Then v satisfies (3.7). Using the well-known asymptotic expansions of the modified Bessel functions or the Liouville-Green approximations (see [7]), we easily obtain

$$w(\eta) \sim \begin{cases} c_1 \eta^{-m/4} \exp\left(-\dfrac{2}{m+2}\eta^{(m+2)/2}\right) & \text{as } \eta \to \infty, \\ c_2(-\eta)^{-m/4} \exp\left(\dfrac{2}{m+2}\eta^{(m+2)/2}\right) & \text{as } \eta \to -\infty \end{cases} \tag{3.10}$$

for some constants c_1, c_2. It is then straightforward to verify (3.6) by using (3.9) and (3.7).

A formal composite approximation can be defined by

$$a(x, \xi, \varepsilon) = u_0(x) + \varepsilon^{4/(m+2)} \frac{F(x)}{F(0)} \frac{f_u(0, u_0(0))}{f_u(x, u_0(x))} v(\xi). \tag{3.11}$$

We show in the Appendix that for any constant $\delta > 0$,

$$\varepsilon^2 a'' - x^m f(x, a) - \varepsilon^2 g(x, a, \varepsilon) = \begin{cases} \mathcal{O}(\varepsilon^4), & \text{for } \delta \leqq |x| \leqq 1, \\ \mathcal{O}(\varepsilon^{4((m+3)/(2m+5))}) & \text{for } |x| \leqq 1. \end{cases} \tag{3.12}$$

For large $|\xi|$, we have

$$a(x, \varepsilon) = u_0(x) + \frac{\varepsilon^2 F(x)}{x^m f_u(x, u_0(x))} + \mathcal{O}\left(\frac{1}{\xi^2}\right).$$

Define

$$\bar{a}(x)=\begin{cases} a(x) & \text{if } |x|\leqq .5, \\ u_0(x)+\dfrac{\varepsilon^2 F(x)}{x^m f_u(x, u_0(x))}+\Lambda(x)\mathcal{O}\left(\dfrac{1}{\xi^2}\right) & \text{if } |x|>.5, \end{cases} \tag{3.13}$$

where Λ is a cutoff function on the interval $[-1, 1]$ having the value 1 near $x=0$ and having the value zero near $x=-1$ and $x=1$.

Finally, to obtain an approximation that is also valid near the endpoints, we require boundary layer corrections of the form

$$B_1\left(\frac{1-x}{\varepsilon}, \varepsilon\right)=B_1\left(\frac{1-x}{\varepsilon}, 0\right)+\mathcal{O}(\varepsilon),$$

$$B_{-1}\left(\frac{x+1}{\varepsilon}, \varepsilon\right)=B_{-1}\left(\frac{x+1}{\varepsilon}, 0\right)+\mathcal{O}(\varepsilon),$$

at the endpoints $x=1$ and $x=-1$, respectively. Since the construction of these corrections has been given by many authors (see [2], [3], [9]), we refer the reader to these sources for the details. The global approximation is

$$A=\bar{a}+\Gamma B_{-1}+(1-\Gamma)B_1 \tag{3.14}$$

where Γ is a cutoff function that is 1 near $x=-1$ and is zero near $x=1$. Then A satisfies (3.12) as well as (3.2).

The next theorem will demonstrate that A gives a uniform approximation of a solution u of (3.1), (3.2) for sufficiently small $\varepsilon>0$. Note that since $A-u_0(x)=\mathcal{O}(\varepsilon^{4/(m+2)})$ near $x=0$, the theorem implies that the graph of $u-u_0(x)$ has a small "bump" as a result of the singularity at $x=0$ and the size of the bump increases with m.

THEOREM 2. *Assume hypotheses* (a)-(e). *For* $\varepsilon>0$ *sufficiently small,* (3.1), (3.2) *has a locally unique solution* $u(x, \varepsilon)$ *so that*

$$u(x, \varepsilon)-A(x, \varepsilon)=\mathcal{O}(\varepsilon^{(10m+24/(2m+5)(m+2))})$$

uniformly in $[0, 1]$, *where* A *is defined by* (3.14), (3.13), (3.9), *and* (3.8).

Proof. The key element in the proof is the construction of an appropriate approximate Green's function for (3.1), (3.2). Let $\lambda(t)=f_u^{1/2}(t, u_0(t))$ for $-1\leqq t\leqq 1$. We will define $H_m(x, t, \varepsilon)$ to be the solution of the problem:

$$\begin{aligned} &\varepsilon^2\frac{d^2H_m}{dx^2}-\lambda^2 x^m H_m=\delta(t-x), \\ &H_m(-1, t)=H_m(1, t)=0. \end{aligned} \tag{3.15}$$

Now $w(\pm(\lambda/\varepsilon)^{2/(m+2)}x)$ are independent solutions of the homogeneous equation. Define

$$h_1\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right)=w\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right)-\frac{w((\lambda/\varepsilon)^{2/(m+2)})}{w(-(\lambda/\varepsilon)^{2/(m+2)})}w\left(-\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right),$$

$$h_2\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right)=w\left(-\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right)-\frac{w((\lambda/\varepsilon)^{2/(m+2)})}{w(-(\lambda/\varepsilon)^{2/(m+2)})}w\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right),$$

so that h_1, h_2 satisfy the homogeneous equation, h_1 vanishes at $x=1$, h_2 vanishes at $x=-1$, and

$$(3.16)\quad h_2(-\eta),\, h_1(\eta)=\begin{cases}\mathcal{O}\left(\eta^{-m/4}\exp\left(-\dfrac{2}{m+2}\,\eta^{(m+2)/2}\right)\right) & \text{as } \eta\to\infty,\\ \mathcal{O}\left((-\eta)^{-m/4}\exp\left(\dfrac{2}{m+2}\,\eta^{(m+2)/2}\right)\right) & \text{as } \eta\to-\infty.\end{cases}$$

Then specifically we have

$$(3.17)\quad H_m(x,t,\varepsilon)=\begin{cases}-C\varepsilon^{-(2m+2)/(m+2)}h_1\left(\left(\dfrac{\lambda}{\varepsilon}\right)^{2/(m+2)}t\right)h_2\left(\left(\dfrac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right) & \text{if } -1\leqq x\leqq t\leqq 1,\\ -C\varepsilon^{-(2m+2)/(m+2)}h_1\left(\left(\dfrac{\lambda}{\varepsilon}\right)^{2/(m+2)}x\right)h_2\left(\left(\dfrac{\lambda}{\varepsilon}\right)^{2/(m+2)}t\right) & \text{if } -1\leqq t\leqq x\leqq 1,\end{cases}$$

where C is a constant independent of ε.

Let L_ε be the linearization of (3.1) about A (see (2.3)). Then

$$\begin{aligned}L_\varepsilon H_m &= \varepsilon^2\frac{d^2H_m}{dx^2}-x^m f_u(x,A)H_m+\mathcal{O}(\varepsilon^2)\\ (3.18)\qquad &= \varepsilon^2\frac{d^2H_m}{dx^2}-\lambda^2x^mH_m+[\lambda^2-f_u(x,A)]x^mH_m+\mathcal{O}(\varepsilon^2)\\ &= \delta(t-x)+\{\mathcal{O}(|t-x|)-f_{uu}(x,\sim)[\mathcal{O}(\varepsilon^{4/(m+2)}v)+B_{-1}+B_1]\}x^mH_m+\mathcal{O}(\varepsilon^2),\end{aligned}$$

where $\sim$ is between u_0 and A.

In a similar way, $H_0(x,t,\varepsilon)$ is defined to be the solution of

$$\varepsilon^2\frac{d^2H_0}{dx^2}-\lambda^2t^mH_0=\delta(t-x),$$

$$H_0(-1,t)=H_0(1,t)=0.$$

Note that the explicit formula for H_0 involves only exponential functions. Now

$$\begin{aligned}L_\varepsilon H_0 &= \varepsilon^2\frac{d^2H_0}{dx^2}-x^mf_u(x,A)H_0+\mathcal{O}(\varepsilon^2)\\ (3.19)\qquad &= \varepsilon^2\frac{d^2H_0}{dx^2}-\lambda^2t^mH_0+[\lambda^2t^m-f_u(x,A)x^m]H_0+\mathcal{O}(\varepsilon^2)\\ &= \delta(t-x)+\mathcal{O}(|x-t|)H_0-f_{uu}(\varepsilon^{4/(m+2)}v+B_{-1}+B_1)x^mH_0+\mathcal{O}(\varepsilon^2).\end{aligned}$$

According to the calculations in [2], we can choose M_{-1}, M_1 to be solutions of

$$(3.20)\quad \varepsilon^2\frac{d^2M_{-1}}{dx^2}-f_u\left(-1,u_0(-1)+B_{-1}\left(\frac{x+1}{\varepsilon},0\right)\right)M_{-1}=f_{uu}B_{-1}\left(\frac{x+1}{\varepsilon},0\right)H_0,$$

$$(3.21)\quad \varepsilon^2\frac{d^2M_1}{dx^2}-f_u\left(1,u_0(1)+B_1\left(\frac{1-x}{\varepsilon},0\right)\right)M_1=f_{uu}B_1\left(\frac{1-x}{\varepsilon},0\right)H_0,$$

which vanish at $x=-1$, $x=1$, respectively, and satisfy

$$M_{-1}=\mathcal{O}\left(\varepsilon^{-1}\exp\left(\frac{-x-1}{\varepsilon}\right)\right),\qquad M_1=\mathcal{O}\left(\varepsilon^{-1}\exp\left(\frac{x-1}{\varepsilon}\right)\right).$$

For M_1 we have

$$\begin{aligned} L_\varepsilon M_1 &= \varepsilon^2 M_1 - f_u(x, A)M_1 + \mathcal{O}(\varepsilon^2) \\ &= f_{uu}B_1\left(\frac{1-x}{\varepsilon}, 0\right)H_0 + \mathcal{O}(|x-1|M_1) + \mathcal{O}(\varepsilon M_1) + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{3.22}$$

A similar equation holds for M_{-1}.

The approximate Green's function is defined as follows:

$$\begin{aligned} \mathrm{Gr}\,(x, t, \varepsilon) &= \Lambda(t)H_m(x, t, \varepsilon) + (1-\Lambda(t))H_0 + \Gamma(x)\chi\left(\frac{t+1}{\varepsilon^\alpha}\right)M_{-1} \\ &\quad + (1-\Gamma(x))\chi\left(\frac{1-t}{\varepsilon^\alpha}\right)M_1, \end{aligned} \tag{3.23}$$

where χ is a cutoff function defined on $[0, \infty]$ which is 1 near zero and is zero when its argument is greater than 1, and $0 < \alpha < 2/(m+2)$.

Among the conditions for a function to be an approximate Green's function, only the third condition is not immediate. Using (3.18), (3.19), and (3.22), we have

$$\begin{aligned} L_\varepsilon\,\mathrm{Gr} - \delta(x-t) &= \Lambda(t)[\mathcal{O}(|t-x|) - f_{uu}(\mathcal{O}(\varepsilon^{4/(m+2)}v) + B_{-1} + B_1)]x^m H_m \\ &\quad + (1-\Lambda(t))[\mathcal{O}(|t-x|) - f_{uu}(\mathcal{O}(\varepsilon^{4/(m+2)}v) + \mathcal{O}(\varepsilon))x^m \\ &\quad + \mathcal{O}((1-x)B_1) + \mathcal{O}((x+1)B_{-1})]H_0 \\ &\quad + [(1-\Gamma(x))\chi((1-t)/\varepsilon^\alpha) \\ &\quad - (1-\Lambda(t))]f_{uu}B_1((1-x)/\varepsilon, 0)H_0 \\ &\quad + [\Gamma(x)\chi((1+t)/\varepsilon^\alpha) \\ &\quad - (1-\Lambda(t))]f_{uu}B_{-1}((1+x)/\varepsilon, 0)H_0 \\ &\quad + (1-\Gamma(x))\chi((1-t)/\varepsilon^\alpha)[\mathcal{O}(1-x) + \mathcal{O}(\varepsilon)]M_1 \\ &\quad + \Gamma(x)\chi((1-t)/\varepsilon^\alpha)[\mathcal{O}(x+1) + \mathcal{O}(\varepsilon)]M_{-1} + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{3.24}$$

The following estimates are valid uniformly for $x \in [-1, 1]$ and for all nonnegative even integers m:

$$\int_{-1}^{1} |t-x||x^m H_m|\, dt = \mathcal{O}(\varepsilon^{2/(m+2)}), \tag{3.25}$$

$$\int_{-1}^{1} |H_m|\, dt = \mathcal{O}(\varepsilon^{-2m/(m+2)}), \tag{3.26}$$

$$\int_{-1}^{1} |vx^m H_m|\, dt = \mathcal{O}(1), \tag{3.27}$$

$$\int_{0}^{1-C\varepsilon^\alpha} \left| B_1\left(\frac{1-x}{\varepsilon}, 0\right)H_0 \right| dt = \text{transcendentally small}. \tag{3.28}$$

We prove estimate (3.25) in Appendix B. The others can be established more easily. It follows from these estimates that Gr is an approximate Green's function.

Finally, from (3.26) we obtain

$$\int_{-1}^{1} |\mathrm{Gr}|\, dt = \mathcal{O}(\varepsilon^{-2m/(m+2)})$$

uniformly for $x \in [-1, 1]$, so

$$|L_\varepsilon^{-1}| = \mathcal{O}(\varepsilon^{-2m/(m+2)}).$$

The result then follows immediately from the corollary to Theorem 1.

Remark. We can show by comparison techniques that $u(x, \varepsilon) - A(x, \varepsilon) = \mathcal{O}(\varepsilon^4)$ uniformly for $0 < \delta \leqq |x| \leqq 1$, where δ is independent of ε and ε is sufficiently small. We omit the details.

4. Problems with a singularity at the boundary. In this final section, we treat the boundary value problem

$$\varepsilon^2 u'' = x^m f(x, u) + \varepsilon^2 g(x, u, \varepsilon), \tag{4.1}$$

$$u(0) = u(1) = 0, \tag{4.2}$$

where m is a positive integer and f satisfies the hypotheses of § 3 with the exception that the inequality in (c) is replaced by

(c)′ $\int_{u_0(0)}^{\theta} f(0, s)\, ds > 0$ for θ between $u_0(0)$ and 0.

Once again, we anticipate boundary layer behavior at the endpoints, but the layer at $x = 0$ will be thicker than the one at $x = 1$. The lowest-order approximation will consist of the reduced solution u_0 plus boundary layer corrections.

With $\xi = x\varepsilon^{-2/(m+2)}$, we find that the first term in the boundary layer correction at zero, $\Theta(\xi)$, satisfies

$$\frac{d^2\Theta}{d\xi^2} = \xi^m f(0, u_0(0) + \Theta), \tag{4.3}$$

$$\Theta(0) = -u_0(0), \quad \Theta(\xi) \to 0 \quad \text{as } \xi \to \infty. \tag{4.4}$$

Let $\lambda(t) = f_u^{1/2}(t, u_0(t))$.

LEMMA. *The problem* (4.3), (4.4) *has a monotone solution* $\Theta(\xi)$ *for which, given* $\delta > 0$, *there is a* $C > 0$ *so that*

$$|\Theta(\xi)| \leqq C \exp\left(\frac{-2\lambda(0) + \delta}{m+2}\, \xi^{(m+2)/2}\right) \tag{4.5}$$

for $\xi \geqq 0$.

Proof. Suppose that $u_0(0) < 0$; the other case is similar. We define a positive, monotone decreasing function $\beta(\xi)$, $\xi \geqq 0$ by

$$\frac{2}{m+2}\, \xi^{(m+2)/2} = \int_\beta^{-u_0(0)} d\tau \Big/ \sqrt{2 \int_0^\tau f(0, u_0(0) + s)\, ds}. \tag{4.6}$$

Note that $\beta(0) = -u_0(0)$,

$$\frac{d\beta}{d\xi} = -\xi^{m/2} \sqrt{2 \int_0^\beta f(0, u_0(0) + s)\, ds} \tag{4.7}$$

and

$$\frac{d^2\beta}{d\xi^2} = \xi^m f(0, u_0(0) + \beta) - \frac{m}{2}\, \xi^{(m-2)/2} \sqrt{2 \int_0^\beta f(0, u_0(0) + s)\, ds}. \tag{4.8}$$

Furthermore, we can use an argument like the one given by Fife [2] for the case $m = 0$ to show that for each $\delta > 0$ there is a $C > 0$ so that

$$\beta(\xi) \leqq c \exp\left(\frac{-2\lambda(0) + \delta}{m+2}\, \xi^{(m+2)/2}\right)$$

for $\xi \geqq 0$.

Now (4.8) implies that β is an upper solution for (4.3), (4.4). In the case $m=1$, β'' is singular at $\xi=0$, but singularities of this type at the boundary are permissible. Since zero is a lower solution, it follows that (4.3), (4.4) has a solution $\Theta(\xi)$ satisfying (4.5).

It remains to show that Θ is monotone decreasing. From (4.3), (4.5), it is clear that $\Theta'(\xi)\to 0$ exponentially fast as $\xi\to\infty$. Multiplying (4.3) by Θ', integrating both sides from ∞ to ξ and using integration by parts, we obtain

$$\begin{aligned}.5\Theta'^2(\xi) &= \int_\infty^\xi \Theta'\Theta''\,d\tau \\ &= \int_\infty^\xi \tau^m f(0, u_0(0)+\Theta)\Theta'\,d\tau \\ &= \xi^m \int_0^{\Theta(\xi)} f(0, u_0(0)+s)\,ds + \int_\xi^\infty m\tau^{m-1}\int_0^{\Theta(\tau)} f(0, u_0(0)+s)\,ds\,d\tau,\end{aligned}$$

so hypothesis (c)′ implies that $\Theta'(\xi)\neq 0$ for all $\xi\geqq 0$, and the proof is complete.

The next theorem yields the existence of a solution having a boundary layer of width $\varepsilon^{2/(m+2)}$ at $x=0$ and a boundary layer of width ε at $x=1$.

THEOREM 3. *Assume hypotheses* (a), (b), (c)′, *and* (d) *are satisfied for the interval* $[0, 1]$. *For all* $\delta>0$ *and sufficiently small* $\varepsilon>0$, (4.1), (4.2) *has a solution* $u(x,\varepsilon)$ *so that* $u(x,\varepsilon)-u_0(x,\varepsilon)=\mathcal{O}(\varepsilon^2)$ *uniformly for* $\delta\leqq x\leqq 1-\delta$.

Proof. We consider the case that $u_0(0)$ and $u_0(1)$ are negative. Existence will follow from the construction of upper and lower solutions for (4.1), (4.2).

An upper solution can be chosen to have the form

$$u_0(x)+\beta(\xi)+\gamma\left(\frac{1-x}{\varepsilon}\right)+C\varepsilon^{4/(m+2)}V(\xi), \tag{4.9}$$

where β is defined by (4.6), γ is a certain positive solution of

$$\varepsilon^2\frac{d^2\gamma}{dx^2}-f(1, u_0(1)+\gamma)<0, \qquad \gamma(0)=-u_0(1),$$

so that $\gamma\to 0$ as $\varepsilon\to 0$ for $x<1$, and C is a positive constant to be determined. The definition of V is similar to the definition of v in § 3. Let $\bar\eta=\xi k^{2/(m+2)}$ and let $w(\bar\eta)$ be defined as in (3.8). Define

$$\bar w(\bar\eta)=\bar\eta^{1/2}I_{1/(m+2)}\left(\frac{2\bar\eta^{(m+2)/2}}{m+2}\right), \tag{4.10}$$

where I is a modified Bessel function, and

$$V(\bar\eta)=Wr^{-1}\left\{\bar w(\bar\eta)\int_{\bar\eta}^\infty w(t)\,dt + w(\bar\eta)\int_0^{\bar\eta}\bar w(t)\,dt\right\},$$

where Wr is the Wronskian. Then V satisfies the equation

$$\frac{d^2V}{d\bar\eta^2}-\bar\eta^m V=-1 \tag{4.11}$$

and

$$V(\bar\eta)\sim\bar\eta^{-m}$$

as $\bar\eta\to\infty$.

It suffices to verify that (4.9) satisfies the requisite differential inequality near $x=0$; see Howes [6, Thm. 4.1] for the verification near $x=1$.

Substituting (4.9) into (4.1), using a prime to denote differentiation with respect to x, and neglecting γ, which is exponentially small, we have

$$\begin{aligned} & x^m f(x, u_0+\beta+C\varepsilon^{4/(m+2)}V(\xi))-\varepsilon^2(u_0''+\beta''+C\varepsilon^{4/(m+2)}V'')+\mathcal{O}(\varepsilon^2) \\ & \quad = x^m[f(x, u_0(x)+\beta)-f(0, u_0(0)+\beta)]+x^m f(0, u_0(0)+\beta)-\varepsilon^2\beta'' \\ & \quad\quad +x^m[f_u(x, \sim)-k^2]C\varepsilon^{4/(m+2)}V+[k^2x^mV-\varepsilon^2V'']C\varepsilon^{4/(m+2)}+\mathcal{O}(\varepsilon^2) \\ & \quad \geqq \mathcal{O}(x^{m+1}\beta)+\frac{m}{2}x^{(m-2)/2}\varepsilon\left(2\int_0^\beta f(0, u_0(0)+s)\,ds\right)^{1/2} \\ & \quad\quad +\mathcal{O}(x^mC\varepsilon^{4/(m+2)}V)+Ck^{4/(m+2)}\varepsilon^2+\mathcal{O}(\varepsilon^2) \end{aligned} \tag{4.12}$$

by the Mean Value Theorem, (4.8), and (4.11). Consider an interval $0 \leqq x \leqq D\varepsilon^{2/(m+2)}$. The second and fourth terms in the last expression in (4.12) are positive and dominate the other terms on this interval for large enough C, so (4.12) is positive on such an interval. Also, we can assume that

$$f_u(x, u+\beta+C\varepsilon^{4/(m+2)}V) \geqq k^2$$

for $x \geqq D\varepsilon^{2/(m+2)}$.

For $x \geqq D\varepsilon^{2/(m+2)}$, the expression (4.12) is at least as large as

$$\begin{aligned} & \mathcal{O}(x^{m+1}\beta)+\frac{m}{2}x^{(m-2)/2}\varepsilon\left(2\int_0^\beta f(0, u_0(0)+s)\,ds\right)^{1/2} \\ & \quad\quad +C\varepsilon^{4/(m+2)}[-\varepsilon^2V''+x^mf_u(x, \sim)V]+\mathcal{O}(\varepsilon^2) \\ & \quad \geqq \mathcal{O}(x^{m+1}\beta)+\frac{m}{2}x^{(m-2)/2}\varepsilon\left(2\int_0^\beta f(0, u_0(0)+s)\,ds\right)^{1/2} \\ & \quad\quad +Ck^{4/(m+2)}\varepsilon^2+\mathcal{O}(\varepsilon^2). \end{aligned}$$

Note that the second term in the last expression dominates the first term for $x \ll \varepsilon^{2/(m+4)}$ and β is exponentially small for $x \gg \varepsilon^{2(m+2)}$. Consequently, (4.12) is positive for all x in a neighborhood of zero if C is large and ε is small.

It is routine to check that $u_0(x)-C\varepsilon^{4/(m+2)}V$ serves as a lower solution for large enough C. These calculations, together with the asymptotic behavior of V, complete the proof for the case considered.

An approximation for a solution of (4.1), (4.2) is defined by

$$A\left(x, \xi, \frac{1-x}{\varepsilon}, \varepsilon\right) = u_0(x)+\Gamma(x)\Theta(\xi)+(1-\Gamma(x))B_1\left(\frac{1-x}{\varepsilon}, \varepsilon\right), \tag{4.13}$$

where B_1 is an appropriate boundary layer correction at $x=1$ and Γ is a cutoff function which has the value 1 near zero and the value zero near 1. A straightforward calculation yields

$$\varepsilon^2A''-x^mf(x, A)-\varepsilon^2g(x, A, \varepsilon) = \mathcal{O}(\varepsilon^{(2m+2)/(m+2)}), \tag{4.14}$$

uniformly for $x \in [0, 1]$ and $A(0)=A(1)=0$. Of course, higher-order approximations would involve terms such as the one in (3.11) to correct the singularity in the regular expansion, as well as additional boundary layer terms.

We can verify the formal approximation (4.13) if hypothesis (c)′ is strengthened.

THEOREM 4. *Assume hypotheses* (a), (b), *and* (d) *are satisfied for the interval* $[0, 1]$, *and*

(c)″ $f(0, s)>0$ *for* s *between zero and* $u_0(0)$.

For all sufficiently small positive ε, (4.1), (4.2) *has a locally unique solution* $u(x, \varepsilon)$ *so that*

$$u(x, \varepsilon) - A(x, \varepsilon) = \mathcal{O}(\varepsilon^{2/(m+2)})$$

uniformly for $0 \leqq x \leqq 1$, *where A is defined by* (4.13), (4.3), *and* (4.4).

Proof. As in the proof of Theorem 2, we let $\lambda(t) = f_u^{1/2}(t, u_0(t))$ and define $H_m(t, x)$ to be the solution of

$$\begin{aligned} &\varepsilon^2 \frac{d^2 H_m}{dx^2} - \lambda^2 x^m H_m = \delta(t-x), \\ &H_m(0, t) = H_m(1, t) = 0. \end{aligned} \tag{4.15}$$

Since m is allowed to be odd, the construction of H_m is slightly different from that given previously. Let w be defined by (3.8), $\bar{w}$ by (4.10), $h_2(s) = \bar{w}(s)$, and

$$h_1(s) = w(s) - \frac{w((\lambda/\varepsilon)^{2/(m+2)})}{h_2((\lambda/\varepsilon)^{2/(m+2)})} h_2(s).$$

Then

$$h_1(s) = \mathcal{O}\left(s^{-m/4} \exp\left(-\frac{2}{m+2} s^{(m+2)/2}\right)\right),$$

$$h_2(s) = \mathcal{O}\left(s^{-m/4} \exp\left(\frac{2}{m+2} s^{(m+2)/2}\right)\right)$$

as $s \to \infty$. With $s = (\lambda/\varepsilon)^{2/(m+2)} x$, h_2, h_2 satisfy the homogeneous portion of (4.11), h_1 vanishes at $x = 1$ and h_2 vanishes at $x = 0$.

Then H_m is defined as in (3.17). Also, H_0 is defined much as in the proof of Theorem 2 with the obvious modifications needed to accommodate the interval [0, 1]. With L_ε as in (2.3),

$$L_\varepsilon H_m = \delta(t-x) + \{\mathcal{O}(|t-x|) - f_{uu}[\Gamma\Theta + (1-\Gamma)B_1]\} x^m H_m + \mathcal{O}(\varepsilon^2), \tag{4.16}$$

$$L_\varepsilon H_0 = \delta(t-x) + \mathcal{O}(|t-x|) H_0 - f_{uu}[\Gamma\Theta + (1-\Gamma)B_1] x^m H_0 + \mathcal{O}(\varepsilon^2). \tag{4.17}$$

The function

$$M_1 = \mathcal{O}(\varepsilon^{-1} \exp(x-1)/\varepsilon)$$

is chosen as in the proof of Theorem 2 to be a solution of (3.21) that vanishes at $x = 1$. M_0 is chosen to be a solution of

$$\frac{d^2 M_0}{d\xi^2} - \xi^m f_u(0, u_0(0) + \Theta) M_0 = f_{uu} \Theta \xi^m H_m, \qquad M_0(0) = 0, \tag{4.18}$$

so that

$$|M_0(\xi)| \leqq C \exp\left(-\frac{2m+2}{m+2}\right) \exp(-\mu \xi^{(m+2)/2}) \tag{4.19}$$

for $\xi \geqq 0$ and some positive constants C and μ. A similar estimate will be valid for M_0'.

Such an M_0 can be constructed in the following way. Note that hypothesis (c)″ implies that $-\theta'$ is a positive upper solution for the homogeneous portion of (4.18). It follows that there is a positive function $z(\xi)$ which satisfies

$$\begin{aligned} &z'' - \xi^m f_u(0, u_0(0) + \Theta) z = 0, \qquad (\xi \geqq 0), \\ &z(0) = -\Theta'(0), \end{aligned}$$

and $z \to 0$ as $\xi \to \infty$. According to the Liouville-Green approximations, for any small $\delta > 0$ there is a $C > 0$ so that

$$C^{-1} \exp\left(-\frac{2}{m+2}(\lambda(0)+\delta)\xi^{(m+2)/2}\right) \leqq z(\xi) \leqq C \exp\left(-\frac{2}{m+2}(\lambda(0)-\delta)\xi^{(m+2)/2}\right)$$

for $\xi \geqq 0$. Define

$$M_0(\xi, t) = z(\xi) \int_0^{\xi} z^{-2}(s) \int_s^{\infty} z(\tau) f_{uu} \tau^m \Theta(\tau) H_m(\tau, t)\, d\tau\, ds.$$

Then M_0 satisfies (4.18) and (4.19).

Now an approximate Green's function is

$$\text{(4.20)} \quad \mathrm{Gr}(x, t, \varepsilon) = \Gamma(t) H_m + (1-\Gamma(t)) H_0 + \Gamma(x)\Gamma\left(\frac{t}{\varepsilon^{\alpha}}\right) M_0 + (1-\Gamma(x))\Gamma\left(\frac{1-t}{\varepsilon^{\alpha}}\right) M_1,$$

where $0 < \alpha < 2/(m+2)$.

The verification that Gr is an approximate Green's function follows from (4.16)-(4.19) and is similar to that given in the proof of Theorem 2. We also have that

$$|L^{-1}| = \mathcal{O}(\varepsilon^{-2m/(m+2)}),$$

which together with the Corollary to Theorem 1 and (4.14) completes the proof of the theorem.

Theorems 2, 3, and 4 contain all the basic ideas needed for investigating a nonoscillatory semilinear singular perturbation problem in which the right-hand side vanishes at one or more points in the interval. In the following example, there are three such points.

Example. Consider the boundary value problem

$$\varepsilon^2 u'' = x^4(1-x^2) f(u) + \varepsilon^2 g(u),$$
$$u(-1) = u(1) = 0.$$

Suppose there is a $C \neq 0$ so that $f(C) = 0$, $f_u(C) > 0$, $\int_C^{\theta} f(u)\, ds > 0$ for θ between C and zero, and $g(C) \neq 0$. Then the hypotheses of Theorems 2 and 3 are satisfied (with appropriate translation of the x-axis) and the boundary value problem has a solution $u(x, \varepsilon)$ if ε is small enough. Furthermore,

$$u(x, \varepsilon) - C = \begin{cases} \mathcal{O}(1) & \text{near } x = -1 \text{ and } x = 1, \\ \mathcal{O}(\varepsilon^{2/3}) & \text{near } x = 0, \\ \mathcal{O}(\varepsilon^2) & \text{elsewhere in } [-1, 1]. \end{cases}$$

The boundary layers at $x = -1$ and $x = 1$ have width $\varepsilon^{2/3}$. We could construct a uniform approximation for $u(x, \varepsilon)$ using the calculations of §§ 3 and 4.

There is a relationship between the problems discussed here and turning point problems for quasilinear equations (see [5], [8]). We will study these latter problems in a future paper.

Appendix A. Verification of (3.12). For the sake of brevity, we write (3.11) as

$$\text{(A1)} \qquad a = u_0 + \varepsilon^{4/(m+2)} b(x) v(\xi).$$

Then

$$\text{(A2)} \qquad \begin{aligned} \varepsilon^2 a'' - x^m f(x, a) - \varepsilon^2 g &= \varepsilon^2(u_0'' - g) + \varepsilon^{2+4/(m+2)}\left(b'' v + \frac{2b'\dot{v}}{\varepsilon^{2/(m+2)}} + \frac{b\ddot{v}}{\varepsilon^{4/(m+2)}}\right) \\ &\quad - x^m[f_u(x, u_0)\varepsilon^{4/(m+2)} b v + \mathcal{O}(\varepsilon^{8/(m+2)} v^2)], \end{aligned}$$

where the prime indicates differentiation with respect to x and the dot indicates differentiation with respect to ξ.

Now the asymptotic expansion for v in negative powers of ξ begins:

$$v(\xi)\sim\frac{(u_0''-g)(0)}{\xi^m f_u(0)}+\frac{m(m+1)(u_0''-g)(0)}{\xi^{2m+2}f_u^2(0)}+\mathcal{O}(\xi^{-3m-4}) \tag{A3}$$

as $\xi\to\pm\infty$. If $|x|\gg\varepsilon^{2/(m+2)}$, (A2) can be written in the form

$$\begin{aligned}&\varepsilon^2(u_0''-g)(x)+\mathcal{O}(\xi^{-m}\varepsilon^{2+(4/(m+2))})+\mathcal{O}(\xi^{-m-1}\varepsilon^{2+(2/(m+2))})\\&\quad+\varepsilon^2 b\frac{(u_0''-g)(0)m(m+1)}{\xi^{m+2}f_u(0)}+\mathcal{O}(\varepsilon^2\xi^{-2m-4})-x^m f_u(x,u_0)\varepsilon^{4/(m+2)}b\\&\quad\cdot\left[\frac{(u_0''-g)(0)}{\xi^m f_u(0)}+\frac{m(m+1)(u_0''-g)(0)}{\xi^{2m+2}f_u^2(0)}\right]\\&\quad+\mathcal{O}(x^m\varepsilon^{4/(m+2)}\xi^{-3m-4})+\mathcal{O}(x^m\varepsilon^{8/(m+2)}\xi^{-2m}),\end{aligned} \tag{A4}$$

where $f_u(0)=f_u(0,u_0(0))$. Using $\xi=x\varepsilon^{-2/(m+2)}$ and

$$b(x)=\frac{u_0''(x)-g(x,u_0(x),0)}{u_0''(0)-g(0,u_0(0),0)}\,\frac{f_u(0,u_0(0))}{f_u(x,u_0(x))},$$

we have

$$(\mathrm{A4})=\mathcal{O}(\varepsilon^4x^{-m-1})+\mathcal{O}(\varepsilon^4x^{-m})+\mathcal{O}(\varepsilon^6x^{-2m-4}). \tag{A5}$$

It follows immediately from (A5) that

$$(\mathrm{A2})=\mathcal{O}(\varepsilon^4)\quad\text{if } |x|\geqq\delta>0,$$

where δ is independent of ε, and

$$(\mathrm{A2})=\mathcal{O}(\varepsilon^{2+2\gamma})+\mathcal{O}(\varepsilon^{(2+(\gamma m+\gamma+2)/(m+2))})\quad\text{if } 1\geqq|x|\geqq M\varepsilon^{(2-\gamma)/(m+2)},$$

where γ and M are any positive ε-independent constants.

Finally, consider the case that $|x|\leqq\varepsilon^{(2-\gamma)/(m+2)}$. We have

$$\begin{aligned}(\mathrm{A2})&=\varepsilon^2(u_0''-g)(x)+\mathcal{O}(\varepsilon^{(2+4/(m+2))}v)+\mathcal{O}(\varepsilon^{2+2/(m+2)}\dot v)+\varepsilon^2 b[\ddot v-\xi^m f_u(0,u_0(0))v]\\&\quad+\mathcal{O}(\varepsilon^2x\xi^m v)+\mathcal{O}(\varepsilon^{8/(m+2)}x^m v^2),\end{aligned}$$

which by (3.5) and the fact that $b(0)=1$ yields

$$\begin{aligned}(\mathrm{A2})&=\mathcal{O}(\varepsilon^2x)+\mathcal{O}(\varepsilon^{(2+4/(m+2))}v)+\mathcal{O}(\varepsilon^{(2+2/(m+2))}\dot v)+\mathcal{O}(\varepsilon^2x\xi^m v)+\mathcal{O}(\varepsilon^{8/(m+2)}x^m v^2)\\&=\mathcal{O}(\varepsilon^{(2m+6-\gamma)/(m+2)}).\end{aligned}$$

From the preceding calculations, it follows that the best estimate is obtained by setting $2\gamma=(2-\gamma)/(m+2)$, so that $\gamma=2/(2m+5)$. The uniform estimate

$$(\mathrm{A2})=\mathcal{O}(\varepsilon^{(4m+12)/(2m+5)})$$

is obtained, which agrees with (3.12).

Appendix B. Verification of (3.25). We will show only that

$$\int_0^x |t-x| x^m H_m(x, t, \varepsilon)\, dt = \mathcal{O}(\varepsilon^{2/(m+2)}) \tag{B1}$$

uniformly for $0 \leqq x \leqq 1$ since estimates over the other t intervals are similar.

Let M be a positive constant and consider first the case $0 \leqq x \leqq \varepsilon^{2/(m+2)}$. Using (3.17), we have that the integral in (B1) is (except for an inessential constant multiple)

$$\begin{aligned} &\varepsilon^{-(2m+2)/(m+2)} x^m h_1\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)} x\right) \int_0^x (x-t) h_2\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)} t\right) dt \\ &= \frac{x^m h_1((\lambda/\varepsilon)^{2/(m+2)} x)}{\lambda^{2/(m+2)} \varepsilon^{(2m-2)/(m+2)}} \int_0^{x\varepsilon^{-2/(m+2)}} \left(\frac{x}{\varepsilon^{2/(m+2)}} - \frac{s}{\lambda^{2/(m+2)}}\right) h_2(s)\, ds \qquad \text{(B2)} \\ &= \mathcal{O}(\varepsilon^{2/(m+2)}). \end{aligned}$$

Next, for $x \geqq M\varepsilon^{2/(m+2)}$, we have

$$\begin{aligned} &\frac{x^m h_1((\lambda/\varepsilon)^{2/(m+2)} x)}{\varepsilon^{(2m+2)/(m+2)}} \int_0^{M\varepsilon^{2/(m+2)}} (x-t) h_2\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)} t\right) dt \\ &= \frac{x^{m+1} h_1((\lambda/\varepsilon)^{2/(m+2)} x)}{\lambda^{2/(m+2)} \varepsilon^{2m/(m+2)}} \left[\int_0^M h_2(s)\, ds + \mathcal{O}(1)\right] \qquad \text{(B3)} \\ &= \mathcal{O}(\varepsilon^{2/(m+2)}). \end{aligned}$$

Finally, from (3.16) we have

$$\begin{aligned} &\frac{x^m h_1((\lambda/\varepsilon)^{2/(m+2)} x)}{\varepsilon^{(2m+2)/(m+2)}} \int_{M\varepsilon^{2/(m+2)}}^x (x-t) h_2\left(\left(\frac{\lambda}{\varepsilon}\right)^{2/(m+2)} t\right) dt \\ &\leqq \frac{x^m}{\varepsilon^{(2m+2)/(m+2)}} \int_{M\varepsilon^{2/(m+2)}}^x (x-t) \\ &\quad \cdot \left(\exp\left(-\frac{2}{\varepsilon(m+2)} x^{(m+2)/2}\right) \Big/ (x\varepsilon^{-(2/(m+2))})^{m/4}\right) \qquad \text{(B4)} \\ &\quad \cdot \left(\exp\left(\frac{2}{\varepsilon(m+2)} t^{(m+2)/2}\right) \Big/ (t\varepsilon^{-(2/(m+2))})^{m/4}\right) dt \\ &= \frac{x^{3m/4}}{\varepsilon} \int_{M\varepsilon^{2/(m+2)}}^x \frac{(x-t)}{t^{m/4}} \exp\left(\frac{2}{\varepsilon(m+2)} (t^{(m+2)/2} - x^{(m+2)/2})\right) dt \\ &\leqq C \frac{x^{3m/4}}{\varepsilon} \int_{M\varepsilon^{2/(m+2)}}^x \frac{(x-t)}{t^{m/4}} \exp\left(\frac{2(t-x)}{\varepsilon(m+2)} x^{m/2}\right) dt \end{aligned}$$

since $t^{(m+2)/2} - x^{(m+2)/2} \leqq (t-x)x^{m/2}$ for $0 \leqq t \leqq x$. Using integration by parts on the last expression in (B4), we have that the first expression in (B4) is no greater than

$$\begin{aligned} C \frac{x^{3m/4}}{\varepsilon^{(3m+4)/(2m+4)}} \Bigg[&(M\varepsilon^{2/(m+2)} - x) \frac{\varepsilon(m+2)}{2x^{m/2}} \exp\left(\frac{2}{\varepsilon(m+2)} (M\varepsilon^{2/(m+2)} - x) x^{m/2}\right) \\ &+ \frac{\varepsilon^2 (m+2)^2}{4x^m} \left(1 - \exp\left(\frac{2}{\varepsilon(m+2)} (M\varepsilon^{2/(m+2)} - x) x^{m/2}\right)\right)\Bigg] \qquad \text{(B5)} \\ = \mathcal{O}(\varepsilon^{2/(m+2)+\delta m/4}) \end{aligned}$$

if $x = \mathcal{O}(\varepsilon^{2/(m+2)-\delta})$ and $0 < \delta < 2/(m+2)$.

The estimates (B2), (B3), and (B5) imply that (B1) is valid uniformly for $x \in [0, 1]$.

REFERENCES

[1] J. A. COCHRAN, *On the uniqueness of solutions of linear differential equations*, J. Math. Anal. Appl., 22 (1968), pp. 418–426.

[2] P. C. FIFE, *Semilinear elliptic boundary value problems with small parameters*, Arch. Rational Mech. Anal., 52 (1973), pp. 205–232.

[3] A. VAN HARTEN, *Nonlinear singular perturbation problems: Proofs of correctness of a formal approximation based on a contraction mapping principle in a Banach space*, J. Math. Anal. Appl., 65 (1978), pp. 126–168.

[4] A. VAN HARTEN AND E. VADER-BURGER, *Approximate Green functions as a tool to prove correctness of a formal approximation in a model of competing and diffusing species*, Pacific J. Math., 125 (1986), pp. 225–249.

[5] F. A. HOWES, *Singularly perturbed nonlinear boundary value problems with turning points*, SIAM J. Math. Anal., 6 (1975), pp. 644–660.

[6] ———, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Memoirs American Mathematical Society 203, Providence, RI, 1978.

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

[8] R. E. O'MALLEY, JR., *On boundary value problems for a singularly perturbed equation with a turning point*, SIAM J. Math. Anal., 1 (1970), pp. 479–490.

[9] D. R. SMITH, *Singular-Perturbation Theory*, Cambridge University Press, Cambridge, 1985.

[10] W. WASOW, *Linear Turning Point Theory*, Lecture Notes in Applied Mathematical Sciences 54, Springer-Verlag, Berlin, New York, 1985.

[11] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 494-503, March 1989

016

ON THE EXISTENCE OF SOLUTIONS OF A TWO-POINT BOUNDARY VALUE PROBLEM ARISING FROM FLOWS IN A CYLINDRICAL FLOATING ZONE*

CHUNQUING LU† AND NICHOLAS D. KAZARINOFF‡

Abstract. Existence of one solution for a two-point boundary value problem arising from flows in a cylindrical floating zone is proved using forward and backward shooting and the Schauder fixed point theorem. If $Q = 2A^3$ Re, where A is the aspect ratio and Re is the Reynolds number, then the existence theorem holds only for $Q < 1$.

Key words. nonautonomous two-point boundary value problem, nonlinear backward shooting, Schauder fixed point theorem

AMS(MOS) subject classification. 34

1. Introduction. We consider solutions of the following nonautonomous two-point boundary value problem (TPBVP) on $[0, 1]$:

$$\begin{aligned} &\text{(a)} \quad [\eta(f'/\eta)']' + Q[f(f'/\eta)' - \eta(f'/\eta)^2] = \beta\eta \qquad (' = d/d\eta), \\ &\text{(b)} \quad f(0) = f(1) = 0 \quad \text{and} \quad (f'/\eta)'|_{\eta=0} = (f'/\eta)'|_{\eta=1} - 1 = 0. \end{aligned} \tag{1.1}$$

This problem arises in the study of surface-tension induced flows of a liquid metal or semiconductor in a cylindrical floating zone of length $2L$ and radius R. In dimensionless coordinates (r, x), points of the cylinder are given by $-1 \leqq x = X/L \leqq 1$, $0 \leqq \eta = r/R \leqq 1$, with free surface $\eta = 1$. The (x, r)-components of dimensionless velocity (u, v) are, respectively, $u = 2A^3(\text{Re})f/\eta$ and $v = -2A^3(\text{Re})f'/\eta$, where Re is the Reynolds number ($Q = 2A^3$ Re), and $A = L/R$ is the aspect ratio. Assuming that the dimensionless pressure p is a quadratic function of x, we find that the r-component of the acceleration equation in the Navier–Stokes energy system describing the flow of fluid and its temperature in the cylinder becomes (1.1)(a). The physical boundary conditions reduce to the conditions (1.1)(b) if we make the assumption that the free boundary is time-independent but not "flat."

Numerical solutions of (1.1) have been found [1] for $0 \leqq Q \leqq 32.7$ and $Q \geqq 1749$ (see Fig. 1). At $Q = 0$, the unique solution of (1.1) is $\eta^2(\eta^2 - 1)/8$. We prove existence of at least one solution of (1.1) for $Q \in [0, Q_0)$, with Q_0 sufficiently small by applying the Schauder fixed point theorem (see [2], where Lu et al. have used the same method to study a simpler TPBVP). The nonautonomous nature of (1.1) makes the proof of our main theorem longer, more complex, and more delicate than the proof of the corresponding result, Theorem 2 of [2]. We are also able to give a crude estimate of how small Q must be for our existence theorem to hold, namely, $Q < Q_0 = 1.0$.

The solutions we have found all correspond to two-cell solutions on the left-hand branch in Fig. 1; they all have $f(\eta) < 0$ on $(0, 1)$ with f' vanishing just once there. Thus u preserves its sign and v vanishes exactly once on $(0, 1)$, which means that these solutions give rise to but two flow cells, one in each half ($x > 0$ and $x < 0$) of the cylinder. For some of the numerically found solutions on the right-hand branch, f does change sign once on $(0, 1)$. These solutions correspond to three-cell flows in the

* Received by the editors August 3, 1987; accepted for publication (in revised form) May 26, 1988.
† Institute of Software, Academica Sinica, P.O. Box 8718, Beijing, People's Republic of China.
‡ Department of Mathematics, State University of New York, Buffalo, New York 14214-3093.

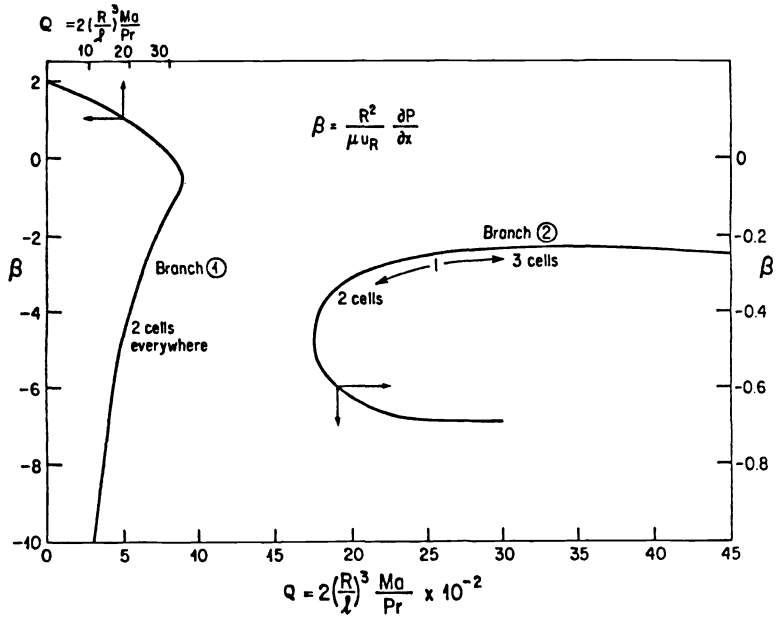


FIG. 1

cylinder. The mathematical existence of such solutions, as well as the existence of others, is still an open question.

2. A nonlinear operator T. Let $g=(f'/\eta)'$. Then (1.1)(a) can be rewritten as

$$g'+[(1+Qf)/\eta]g-Q(f'/\eta)^2=\beta. \tag{2.1}$$

We differentiate (1.2) once with respect to η and obtain the equation

$$g''+[(1+Qf)/\eta]g'-[\{1+Q(\eta f)'\}/\eta^2]g=0 \qquad (0<\eta<1) \tag{2.2}$$

to which we add the conditions

$$\lim_{\eta\searrow 0} g(\eta)=g(0)=g(1)-1=0. \tag{2.3}$$

Let

$$\Omega=\{f\,|\,f\in C^1[0,1], f(0)=0, f'(0)=0, f(1)=0, -\tfrac{1}{3}\leqq f(\eta)\leqq 0,$$
$$-\tfrac{2}{3}\leqq f'(\eta)\leqq 1, \text{ and } \|f\|\leqq\tfrac{4}{3}\}$$

where $\|f\|=\max|f|+\max|f'|$, and the maxima are taken over $[0,1]$. It is easy to check that Ω is a nonempty, closed, bounded, convex subset of $C^1[0,1]$.

Given any $f\in\Omega$, find the solution g of the linear TPBVP (2.2)-(2.3), and let f^* be the solution of the TPBVP $(f^{*\prime}/\eta)'=g(\eta\in(0,1))$ with $f^*(0)=f^*(1)=0$. Then define $Tf=f^*$. Thus, given a $Q\in[0,Q_0]$ for some $Q_0>0$, if we can find an $f\in\Omega$ for which $Tf=f$, then f will be a solution of (1.1) with

$$f(\eta)=\int_0^\eta\left[\int_0^s g(t)\,dt\right]ds+\frac{1}{2}\eta^2 f''(0) \quad\text{and}\quad f''(0)=-2\int_0^1\left[s\int_0^s g(t)\,dt\right]ds.$$

We carry out the above program for defining f^* and $Tf=f^*$ in the next section. In § 4 we show that $T(\Omega)\subseteq\Omega$, T is continuous and T maps bounded subsets of Ω

into compact subsets of Ω. Having thus set the stage, in § 5, we apply Schauder's fixed point theorem [3] and obtain our main theorem.

3. Some lemmas. We now consider the following backward initial value problem on $(0, 1)$:

$$g''+[(1+Qf)/\eta]g'-[\{1+Q(\eta f)'\}/\eta^2]g=0 \quad \text{with } g(1)=1 \text{ and } g'(1)=\alpha. \tag{3.1}$$

We multiply (3.1) by the integrating factor $I(\eta)=\exp\{\int_1^\eta [(1+Qf(s))/s]\,ds\}$, integrate the differential equation in (3.1) from 1 to η, and obtain

$$g'(\eta)I(\eta)=\int_1^\eta \{[1+Q(tf)']/t^2\}g(t)I(t)\,dt+\alpha. \tag{3.2}$$

In each of the eight lemmas below the following hypotheses are to be understood: $f\in\Omega$, g is a solution of (3.1), and $Q\in[0, 1)$.

LEMMA 1. *If $\alpha=g'(1)\leqq 0$, then $g'(\eta)\leqq 0$ for all $\eta\in(0, 1)$.*

Proof. Observe that since $f\in Q$,

$$1+Q(tf)'=1+Q(f+tf')>1-Q(\tfrac{1}{3}+2t/3)>0 \tag{3.3}$$

for all $Q\in[0, 1]$ and all $t\in(0, 1)$. If $g'(1)\leqq 0$, then initially for $t<1$, $g(t)\geqq 1$. Hence, the integral on the right-hand side of (3.2) is negative as long as $g(t)>0$, and so, we claim, $g'(\eta)\leqq 0$ for all $\eta\in(0, 1)$. If not, $g'(\eta_0)>0$ for some $\eta_0\in(0, 1)$, and there must be a first zero $\eta_1>\eta_0$ of g' (closest to one) for which $g''(\eta_1)<0$. But from the differential equation (3.1) $g''(\eta_1)>0$. This is a contradiction, and the proof of Lemma 1 is complete. □

LEMMA 2. *If $g(\eta_0)=0$ for some $\eta_0\in(0, 1)$, then $g(\eta)<0$ for $0<\eta<\eta_0$; that is, g has no more than one zero and*

$$\lim_{\eta\searrow 0} g(\eta)<0.$$

Proof. Suppose not. Let $\eta_0<1$ be the zero of g closest to 1 and let η_1 be the next zero of g. Then there must be an $\eta_2\in(\eta_1, \eta_0)$ such that $g(\eta_2)<0$ and $g'(\eta_2)=0$. To see this we note that, since η_1 is the first zero of g less than η_0 and $g(1)=0$, $g(\eta)<0$ on (η_1, η_0), $g(\eta_2)$ is a local minimum of g and $g''(\eta_2)>0$. However, the differential equation in (3.1) implies $g''(\eta_2)<0$ since $g'(\eta_2)=0$, $g(\eta_2)<0$ and $1-Q(tf)'>0$. This is a contradiction.

It remains to prove that $\lim_{\eta\searrow 0} g(\eta)<0$. We first observe that since $g(\eta_0)=0$, $g'(\eta_0)\neq 0$, for otherwise $g\equiv 0$. Then $g'(\eta_0)>0$, and by (3.1), $g''(\eta_0)<0$. Thus $g'(\eta)>g'(\eta_0)$ and $g(\eta)<0$ for $\eta<\eta_0$ and close to η_0. Moreover, $g'(\eta)$ remains positive on $(0, \eta_0)$, because $g''(\eta)<0$ wherever $g'(\eta)=0$ for $\eta<\eta_0$. Therefore, $\lim_{\eta\searrow 0} g(\eta)<0$ (the limit may be $-\infty$). This completes the proof. □

LEMMA 3. *Let $Q\in[0, 1)$ be given, $f\in\Omega$, and $g(\eta, \alpha)$ solve the backward initial value problem* (3.1). *Then there exists an $\alpha_0>0$ such that $\alpha>\alpha_0$ implies an $\eta\in(\frac{3}{4}, 1)$ exists at which $g(\eta, \alpha)=0$.*

Proof. Rewrite the differential equation in (3.1) as follows:

$$g''=-[(1+Qf)/\eta]g'+[\{1+Q(\eta f)'\}/\eta^2]g. \tag{3.4}$$

Let α_1 satisfy the condition

$$-(1-Q/3)\alpha_1+(1+Q\|f\|)/(\tfrac{1}{2})^2<0. \tag{3.5}$$

Then for all $\alpha>\alpha_1$ and $\eta\in[\frac{1}{2}, 1)$,

$$-(1-Q/3)\alpha/\eta+g(1; a)(1+Q\|f\|)/(\eta)^2<0;$$

hence, $g''(1;\alpha)<0$. This implies that for $\eta<1$ and η close to 1, $g'(\eta;\alpha)>g'(1;\alpha)=\alpha$ and $g(\eta;\alpha)<1$. We claim $g''(\eta;\alpha)<0$ as long as $g'(\eta;\alpha)>\alpha$ for $\eta\in[\frac{1}{2},1)$. To see this, we use (3.4)-(3.5) together with $g'(\eta;\alpha)>\alpha$ and $g(\eta;\alpha)<g(1;\alpha)=1$ to obtain the following estimate: for $\alpha>\alpha_1$,

$$\begin{aligned} g''(\eta;\alpha) &< -(1-Q/3)\alpha/\eta+g(1,\alpha)(1+Q\|f\|)/(\eta)^2 \\ &< -(1-Q/3)\alpha_1+(1+Q\|f\|)/(\tfrac{1}{2})^2 \qquad (\eta\in[\tfrac{1}{2},1)) \\ &< 0. \end{aligned} \tag{3.6}$$

Then g' increases as η decreases; hence, $g''(\eta;\alpha)<0$ and $g'(\eta;\alpha)>\alpha$ for $\eta\in[\frac{1}{2},1)$.

Using Taylor's formula, for $\eta\in[\frac{1}{2},1)$ we can write

$$g(\eta;\alpha)=g(1;\alpha)+g'(1;\alpha)(\eta-1)+g''(\xi;\alpha)(\eta-1)^2/2,$$

for some $\xi\in(\eta;1)$. From this and (3.6) we see that if $\alpha_0=\max[\alpha_1,4]$, then $g(\frac{3}{4};\alpha)<0$ for $\alpha>\alpha_0$. Finally, since $g(1;\alpha)=1$, we conclude that g must vanish somewhere on $(\frac{3}{4},1)$. The proof of the lemma is complete. □

LEMMA 4. *There is a nonempty set A such that $A\subset(0,\alpha_0)$ and*

$$0\leqq\lim_{\eta\searrow 0} g(\eta;\alpha)<1 \quad \textit{for } \alpha\in A.$$

Proof. We define two subsets of the real line as follows:

$$A_1=\{\alpha\,|\,g'(\eta;\alpha)\leqq 0 \text{ for some } \eta\in(0,1]\},$$
$$A_2=\{\alpha\,|\,g(\eta;\alpha)=0 \text{ for some } \eta\in(0,1)\}.$$

By Lemma 3, $A_2\neq\varnothing$; and A_2 is open since, by (3.1), $g'(\eta;\alpha)\neq 0$ if $g(\eta;\alpha)=0$. By Lemma 1, $A_1\neq\varnothing$, and A_1 is also open since, by (3.1), $g(\eta;\alpha)\neq 0$ if $g'(\eta;\alpha)=0$. We claim $A_1\cap A_2=\varnothing$. Suppose not. First, we observe that it is impossible for $g'(\eta)<0$ and the $g(\eta)=0$ to hold simultaneously. If $g(\eta)$ vanishes farther from 1 than $g'(\eta)$ becomes negative, then there must be a local maximum of g at a point η_1 where $g'(\eta_1)=0$, $g''(\eta_1)<0$, and $g(\eta_1)>0$. This contradicts (3.1). If $g'(\eta)<0$ farther from one than an $\eta<1$ exists where $g(\eta)=0$, then g must take on a negative local minimum, which also contradicts (3.1).

It is clear from the definitions of A_1 and A_2 that the complement $(A_1\cup A_2)^c$ in $\mathbb{R}^1$ of $A_1\cup A_2$ is not empty and $(A_1\cup A_2)^c\subset(0,\alpha_0)$, where α_0 is defined in the proof of Lemma 3 above. Let $A=(A_1\cup A_2)^c$. Then for $\alpha\in A$, $g'(\eta,\alpha)$ and $g(\eta,\alpha)$ are positive on $(0,1)$. Therefore, $\lim_{\eta\searrow 0} g(\eta)$ exists and $0\leqq\lim_{\eta\searrow 0} g(\eta)<1$. This completes the proof. □

We next prove that if $\alpha\in A$, $\lim_{\eta\searrow 0} g(\eta;\alpha)=0$.

LEMMA 5. *If $\lim_{\eta\searrow 0} g(\eta;\alpha)$ exists and is finite, and, in particular, if $\alpha\in A$, then this limit is zero.*

Proof. Recall from § 2 that for $f\in\Omega$, $f(0)=f'(0)=0$ and $f(\eta)<0$ on $(0,1)$. We begin with (3.2). Suppose $\eta_0>0$ is so small that

$$\tfrac{1}{2}<1+Q(\eta f)' \qquad (0\leqq\eta<\eta_0).$$

Then the integral appearing on the right-hand side of (3.2) can be rewritten as

$$\begin{aligned} \int_1^\eta \{[1+Q(tf)']/t^2\}g(t)I(t)\,dt &\equiv \int_1^\eta F(t)\,dt, \\ &= \int_1^{\eta_0} F(t)\,dt+\int_{\eta_0}^{\eta} F(t)\,dt, \\ &\equiv M(\eta_0)+B(\eta). \end{aligned} \tag{3.7}$$

Clearly, $M(\eta_0)<0$. Since $g(t)>0$ on $(0, 1]$, for $0<\eta<\eta_0$

$$\frac{1}{2}\int_{\eta_0}^{\eta} g(t)I(t)/t^2\,dt \geqq B(\eta) \tag{3.8}$$

where $I(t)$ is defined at the beginning of this section. Note that $f'(0)=0$ and $f(\eta)<0$ on $(0, 1)$ imply that given $\varepsilon>0$ there exists an $\eta_1=\eta_1(\varepsilon)>0$ such that

$$-\varepsilon<f(v)/v<0 \qquad (v\in(0, \eta_1)). \tag{3.9}$$

We rewrite the integral $h(t)$ in the exponent for $I(t)$ as

$$\begin{aligned} h(t) &= \int_1^{\eta_1}\{(1+Qf)/v\}\,dv+\int_{\eta_1}^{t}\{(1+Qf)/v\}\,dv \\ &\equiv N(\eta_1)+\ln t-\ln\eta_1+Q\int_{\eta_1}^{t}\left(\frac{f}{v}\right)dv. \end{aligned}$$

Therefore, for $0<t<\eta_1$

$$-Q\varepsilon(\eta_1-t)<h(t)-[N(\eta_1)+\ln t-\ln\eta_1]<Q\varepsilon(\eta_1-t). \tag{3.10}$$

Let $\eta_2=\min[\eta_0, \eta_1]$. Then (3.8) and (3.10) both hold for $0<t<\eta_2$. Therefore,

$$tN^-(\eta_2, \varepsilon)\,e^{+Q\varepsilon t}<I(t)<tN^+(\eta_2, \varepsilon)\,e^{-Q\varepsilon t} \tag{3.11}$$

where

$$N^{\pm}(\eta, \varepsilon)=(1/\eta)\,e^{N(\eta)(\pm Q\varepsilon\eta)}.$$

It follows that for $0<\eta<\eta_2$ the function $B(\eta)$ defined in (3.8) satisfies the following inequalities:

$$B(\eta)<\frac{1}{2}\int_{\eta_2}^{\eta} t^{-1}g(t)N^+(\eta_2, \varepsilon)\,e^{-Q\varepsilon t}\,dt. \tag{3.12}$$

Finally, from (3.2), (3.7), and (3.12), for $0<\eta<\eta_2$ we obtain the inequality

$$\eta N^-(\eta_2, \varepsilon)g'(\eta)\,e^{+Q\varepsilon\eta}<\frac{1}{2}\int_{\eta_2}^{\eta} t^{-1}g(t)N^+(\eta_2, \varepsilon)\,e^{-Q\varepsilon t}\,dt+M(\eta_2)+g'(1). \tag{3.13}$$

If we let $\eta\searrow 0$ in (3.13) and $\lim_{\eta\searrow 0} g(\eta)=\beta>0$, then the integral (3.13) approaches $-\infty$. Thus there is an $\eta^*>0$ such that $g'(\eta)<0$ for $0<\eta<\eta^*$. This is a contradiction. Consequently, $\lim_{\eta\searrow 0} g(\eta; \alpha)=0$. The proof of the lemma is complete. □

LEMMA 6. *There exists only one* $\alpha\in A$ *such that* $\lim_{\eta\searrow 0} g(\eta; \alpha)=0$.

Proof. We prove this by contradiction. Suppose there exist two solutions $g_1(\eta)=g(\eta; \alpha_1)$ and $g_2(\eta)=g(\eta; \alpha_2)$ that approach zero as $n\searrow 0$. Let $g_1-g_2=h$. Then h satisfies the differential equation

$$h''+[(1+Qf)/\eta]h'-[\{1+Q(\eta f)'\}/\eta^2]h=0 \tag{3.14}$$

with $h(1)=0$, $h'(1)=\alpha_1-\alpha_2\neq 0$. Suppose $\alpha_1<\alpha_2$. Then $h(\eta)>0$ and $h''(\eta)>0$ as long as $h'(\eta)<0$. Suppose that for some first point $\eta_0\in(0, 1)$ (closest to 1) $h'(\eta_0)=0$. Then it must be that $h(\eta_0)>0$. Then, by (3.14) we see that $h''(\eta_0)>0$, which is a contradiction. Therefore, (a) $g_1'(\eta)<g_2'(\eta)$ and (b) $g_1(\eta)>g_2(\eta)$. Integrating (a) from zero to η and using our assumption that $g_1(+0)=g_2(+0)=0$, we find that $g_1(\eta)<g_2(\eta)$, which contradicts (b), and the proof is complete. □

The proof of Lemma 5 implies that not only $\lim_{\eta \searrow 0} g(\eta) = 0$ but that $\eta g'(\eta)$ remains bounded as $\eta \searrow 0$. We now show that, under the hypotheses of Lemma 4, $\lim_{\eta \searrow 0} \eta g'(\eta) = 0$.

LEMMA 7. *Under the hypotheses of Lemma* 4, $\lim_{\eta \searrow 0} \eta g'(\eta) = 0$.

Proof. It is obvious that $\lim_{\eta \searrow 0} g'(\eta) I(\eta) \geqq 0$ exists and

$$0 \leqq \lim_{\eta \searrow 0} g'(\eta) I(\eta) < \alpha.$$

Also, $I(\eta)/\eta$ has a finite limit as $\eta \searrow 0$. Hence, $\lim_{\eta \searrow 0} \eta g'(\eta)$ exists and is nonnegative. Suppose this limit is $\nu > 0$. Then given $\varepsilon > 0$ with $0 < \varepsilon < \nu$, there exists an $\eta^{\#} > 0$ such that for $0 < \eta < \eta^{\#}$,

$$\nu - \varepsilon < \eta g'(\eta) < \nu + \varepsilon.$$

We divide these inequalities by η and integrate from η to $\eta^{\#}$. The result is

$$(\nu - \varepsilon)[\ln(\eta^{\#}/\eta)] < g(\eta^{\#}) - g(\eta) < (\nu + \varepsilon)[\ln(\eta^{\#}/\eta)].$$

These inequalities imply that as $\eta \searrow 0$, $g \to -\infty$, which is a contradiction. Hence, $\lim_{\eta \searrow 0} \eta g'(\eta) = 0$. □

Remark. If $\lim_{\eta \searrow 0} g(\eta)$ exists (including $\pm\infty$) and g' has constant sign near $\eta = 0$, then all integral curves of (3.1) are asymptotic to either $\pm\infty$ as $\eta \searrow 0$, except for one that approaches the origin.

We next obtain bounds for g and g' on $[0, 1]$.

LEMMA 8. *If the hypotheses of Lemma* 4 *are satisfied and* $\alpha \in A$, *then the solution* g *of* (3.1) *satisfies the inequalities*

$$\eta^{1+2Q/3} \leqq g(\eta) \leqq \eta^{1-Q/3}, \tag{3.15}$$

$$(1 - Q/3)\eta^{1+2Q/3} \leqq \eta g'(\eta) \leqq (1 + 2Q/3)\eta^{1-Q/3} \tag{3.16}$$

for all $\eta \in [0, 1]$.

Proof. Rewrite (3.1):

$$(\eta^2 g')' - (\eta g)' - Q(\eta f g)' = -2Q\eta f g'. \tag{3.17}$$

We integrate (3.17) from zero to η and apply Lemmas 5 and 6 ($g(0) = 0$ and $\eta g'(\eta)|_{\eta=0} = 0$) to obtain

$$g' = [(1 + Qf)/\eta]g - (2Q/\eta^2) \int_0^{\eta} tfg'\, dt \qquad (0 < \eta < 1). \tag{3.18}$$

Since $-\frac{1}{3} \leqq f(\eta) \leqq 0$ and $g'(\eta) \geqq 0$, using (3.18), we obtain the inequalities

$$[(1 - \tfrac{1}{3}Q)/\eta]g(\eta) \leqq g'(\eta) \leqq (1/\eta)g(\eta) + [2Q/(3\eta^2)] \int_0^{\eta} tg'\, dt \qquad (0 < \eta < 1). \tag{3.19}$$

We apply integration by parts to the integral in (3.19) to find

$$\int_0^{\eta} tg'\, dt = \eta g(\eta) - \int_0^{\eta} g\, dt \qquad (0 < \eta < 1). \tag{3.20}$$

Finally, combining (3.19) and (3.20) and using the positivity of g for $\alpha \in A$, we obtain the inequalities

$$[(1 - \tfrac{1}{3}Q)/\eta]g(\eta) \leqq g'(\eta) \leqq (1/\eta)g(\eta) + [2Q/(3\eta^2)][\eta g(\eta)] = [(1 + 2Q/3)/\eta]g,$$

namely,

$$(1 - \tfrac{1}{3}Q)/\eta \leqq g'(\eta)/g(\eta) \leqq (1 + 2Q/3)/\eta \qquad (0 < \eta < 1). \tag{3.21}$$

If we integrate (3.21) from η to 1 and exponentiate the result, we obtain (3.15). The combination of (3.15) with (3.21) then yields (3.16), and the lemma is proved. □

4. Definition, continuity, compactness of the operator T. We recall the definitions of Ω and T from § 2. The domain Ω of T consists of the functions in $C^1[0,1]$ for which $f(0)=f'(0)=f(1)=0$, $-\frac{1}{3}\leqq f(\eta)\leqq 0$, $-\frac{2}{3}\leqq f'(\eta)\leqq 1$ and $\|f\|\leqq\frac{4}{3}$. The operator T is defined on Ω by $Tf=f^*$, where f^* satisfies the equation $(f^{*\prime}/\eta)'=g$, subject to the boundary conditions $f^*(0)=f^*(1)=f^{*\prime}(0)=0$, and where g is *the* solution of the backward initial value problem (3.1), found in the last section, with α chosen so that $g(0)=0$ and on $(0,1)$ both $g(\eta)$ and $g'(\eta)$ are positive. We first prove Lemma 9.

LEMMA 9. *If the hypotheses of Lemma* 8 *are satisfied, then the operator T is well defined and maps any bounded subset of Ω into a compact subset of Ω.*

Proof. We first show that for $\|f\|\leqq\frac{4}{3}$, as in the definition of Ω, T maps Ω into itself. We integrate the equation $(f^{*\prime}/\eta)'=g$ from zero to η and use the conditions $f^*(0)=f^{*\prime}(0)=0$ and obtain the results that

$$f^{*\prime}(\eta)=k\eta+\eta\int_0^\eta g(t)\,dt, \tag{4.1}$$

and

$$f^*(\eta)=\frac{1}{2}k\eta^2+\int_0^\eta\left\{x\int_0^x g(t)\,dt\right\}dx, \tag{4.2}$$

where

$$k=-2\int_0^1\left\{x\int_0^x g(t)\,dt\right\}dx=f^{*\prime\prime}(0).$$

Then, $f^*(1)=0$. Differentiating (4.1), we find that

$$f^{*\prime\prime}(\eta)=k+\int_0^\eta g(t)\,dt+\eta g(\eta). \tag{4.3}$$

Since g is positive and strictly increasing on $(0,1)$, $f^{*\prime\prime}$ has at most one zero on $(0,1)$, and hence f^* does not change sign on $(0,1)$. But $f^{*\prime\prime}(0)=k<0$ and $f^*(0)=f^{*\prime}(0)=0$ imply that f^* is negative near zero. Thus, $f^*(\eta)<0$ on $[0,1]$.

We now show that f^* and its derivatives have the bounds necessary for f^* to lie in Ω and for T to have the property that if $K\in\Omega$ is closed and bounded, then the closure, cl $(T(K))$, of $T(K)$ is compact. From (4.3) we see that

$$k<f^{*\prime\prime}(\eta)<\int_0^\eta g(t)\,dt+\eta g(\eta)<2, \tag{4.4}$$

where

$$-k=|k|\leqq 2\int_0^1\left\{x\int_0^x dt\right\}dx=\frac{2}{3},$$

and

$$-\tfrac{2}{3}\leqq k<0. \tag{4.5}$$

Using this result in (4.2), we obtain the inequalities

$$0\geqq f^*(\eta)\geqq\tfrac{1}{2}k\eta^2\geqq k/2\geqq-\tfrac{1}{3}.$$

Thus

$$|f^*(\eta)|\leqq\tfrac{1}{3}, \tag{4.6}$$

and, by (4.4)

$$|f^{*\prime\prime}(\eta)| \leqq 2. \tag{4.7}$$

Finally, using the above results in (4.1) we obtain the inequalities

$$-\tfrac{2}{3} \leqq f^{*\prime}(\eta) \leqq 1. \tag{4.8}$$

From (4.6) and (4.8) it now follows that $\|f^*\| \leqq 1+\frac{1}{3}=\frac{4}{3}$. Therefore, if $0 \leqq Q < 1.0 = Q_0$, then $1 = Q(\eta f)' > 0$, and the hypotheses of the previous lemmas hold. Thus g has the desired properties, T is well defined, and $T(\Omega) \subseteq \Omega$. It also follows that if K is a closed and bounded subset of Ω, then cl $(T(K))$ is compact. Suppose $\{f_i\}$ is a sequence of functions in K with images f_i^* under T. Then $|f_i^{*\prime\prime}| \leqq 2$, $|f_i^{*\prime}| \leqq 1$ and $|f_i^*| \leqq \frac{1}{3}$ on $[0, 1]$ for each i.

Therefore, $\{f_i^*\}$ and $\{f_i^{*\prime}\}$ are equicontinuous on $[0, 1]$. Hence, by the Arzela-Ascoli theorem, there exist a subsequence $\{f_{n(i)}^*\}$ of $\{f_i\}$ and a $g \in$ cl $(T(K))$ such that $\|f_{n(i)}^* - g\| \to 0$ as $i \to \infty$. Thus, cl $(T(K))$ is compact. The lemma is proved. □

Last, we prove Lemma 10.

LEMMA 10. *The operator T is continuous.*

Proof. By (4.2) and the definition of k, it is sufficient to prove that for any given $f_0 \in \Omega$ and any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $\|f - f_0\| < \delta$ and $f \in \Omega$, then $\max_{[0,1]} (|g(\eta) - g_0(\eta)|) < \varepsilon$, where g and g_0 are solutions of (3.1) corresponding to coefficients f and f_0, respectively, that are positive and have positive first derivatives on $(0, 1)$ and that vanish at $\eta = 0$.

Step 1. Since the estimates for g in (3.16) hold for any $f \in \Omega$, we obtain from (3.15) the estimate

$$|g(\eta) - g_0(\eta)| \leqq 2\eta^{1-Q/3} \qquad (0 \leqq \eta \leqq 1). \tag{4.9}$$

Then by (4.9), for $0 \leqq \eta \leqq \varepsilon$,

$$|g(\eta) - g_0(\eta)| \leqq 2\varepsilon^{1-Q/3} \equiv \varepsilon_1. \tag{4.10}$$

Step 2. Let $r(\eta) = g(\eta) - g_0(\eta)$. Then, after performing some manipulation, we find from (3.1) that

$$\begin{aligned} L[r] &\equiv r'' + [(1+Qf_0)/\eta]r' - [(1+Q(\eta f_0)')/\eta^2]r \\ &= -(Q/\eta^2)[(f-f_0)(\eta g' - g) - (f' - f_0')\eta g] \\ &\equiv -F(\eta) \end{aligned} \tag{4.11}$$

with $r(0) = r(1) = 0$. Obviously, (4.10) yields $|r(\varepsilon)| < \varepsilon_1$. We shall consider (4.11) on $[\varepsilon, 1]$ and prove that $|r|$ is uniformly as small as we like on $[\varepsilon, 1]$ if f is sufficiently close to f_0 and $0 \leqq Q < 1$. By Lemma 8, g and g' are bounded independently of $f \in \Omega$. Thus if ε^* is any point of the interval

$$0 < \varepsilon^* < \tfrac{1}{2}(1-Q)\varepsilon_1, \tag{4.12}$$

then, if f is close enough to f_0 and $0 \leqq Q < 1$,

$$(Q/\varepsilon^2)[\max\{|\eta g'| + |\eta g| + |g|\}]\|f - f_0\| \leqq (3+2Q/3)(Q/\varepsilon^2)\|f - f_0\| < \varepsilon^*, \tag{4.13}$$

where the maximum is taken over $[0, 1]$.

We may fix f_0. By (4.13), if r is any solution of (4.11) on $[\varepsilon, 1]$ with $r(1) = 0$ and $|r(\varepsilon)| \leqq \varepsilon_1$,

$$L[r] - \varepsilon^* < L[r] + F(\eta) < L[r] + \varepsilon^* \qquad (\varepsilon \leqq \eta \leqq 1). \tag{4.14}$$

Let $r^{\pm}$ denote the solution of $L[r] \pm \varepsilon^* = 0$ satisfying $r^{\pm}(1) = 0$ and $r^{\pm}(\varepsilon) = r(\varepsilon)$, where $L[r] = -F$ and $r(0) = r(1) = 0$. Then by (4.14) the standard comparison theorem applies to (4.11) and yields

$$r^-(\eta) \leqq r(\eta) \leqq r^+(\eta)$$

for $\eta \in [\varepsilon, 1]$.

Step 3. It remains to prove that both $|r^{\pm}(\eta)|$ are small. Fix r so that $|r(\varepsilon)| < \varepsilon_1$. We shall use (4.12) to show that $|r^{\pm}(\eta)| < \varepsilon_1$ on $[\varepsilon, 1]$. To see this consider the problems

$$L[v] \pm \varepsilon^* = 0 \quad \text{with } v(1) = 0 \text{ and } v(\varepsilon) = r(\varepsilon).$$

These are solved by $r^{\pm}$.

If there exists a point x_+ (x_-) in $(\varepsilon, 1)$ for which $r^+(x_+) > \varepsilon_1$ ($r^-(x_-) > \varepsilon_1$), then there must be a point y_+ (y_-) in $(\varepsilon, 1)$ at which one or both of r^+ and r^- take their maximum value, namely, either $r^{+\prime}(y_+) = 0$, $r^{+\prime\prime}(y_+) < 0$, and $r^+(y_+) > \varepsilon_1$, and/or $r^{-1}(y_-) = 0$, $r^{-\prime\prime}(y_-) < 0$, and $r^-(y_-) > \varepsilon_1$. But we also see from (4.12) that one or both of the following hold:

$$\begin{aligned} r^{\pm\prime\prime}(y_\pm) \pm \varepsilon^* &= \{[1 + Qf_0(y_\pm) + Qf_0'(y_\pm)y_\pm]/y_\pm^2\} r^{\pm}(y_\pm), \\ &> (1 - Q)\varepsilon_1 / y_\pm^2, \\ &> (1 - Q)\varepsilon_1. \end{aligned}$$

Hence,

$$r^{\pm\prime\prime}(y_\pm) > (1 - Q)\varepsilon_1 - (\pm\varepsilon^*) > 0;$$

that is, one or both of $r^{\pm\prime\prime}(y_\pm) > 0$. This is a contradiction. If there exists a point at which r^+ (r^-) is less than $-\varepsilon_1$, we argue similarly, and again we obtain a contradiction.

In summary, for any given $\varepsilon_1 > 0$, if we choose

$$\delta = \varepsilon^* / \{3 + 2Q/3)(Q/\varepsilon^2)\},$$

where ε and ε^* satisfy (4.10) and (4.12), respectively, then for $f \in \Omega$ and $\|f - f_0\| < \delta$, we see that $|g(\eta) - g_0(\eta)| < \varepsilon_1$ for all $\eta \in [0, 1]$. But then, by the estimates (4.1)-(4.3) for f^* in terms of g, it follows that $\|f^* - f_0^*\|$ can be made as small as we please if f is sufficiently close to f_0 and $Q \in [0, 1]$. This proves the continuity of T. □

5. The main theorem. The set Ω, defined in § 2 is a nonempty, closed, bounded, convex subset of $C^1[0, 1]$ with the usual norm ($\|f\|$ is the sum of the maxima over $[0, 1]$ of f and f'). We have shown that the operator $T: \Omega \mapsto \Omega$ is continuous and maps bounded subsets of Ω into compact subsets of Ω; that is, T is compact. Therefore, we may apply the Schauder Fixed Point Theorem to T on Ω and conclude that there exists at least one fixed point $f = Tf \in \Omega$ of T. This fixed point is a desired solution to the TPBVP (1.1a), (1.1b). We state this result as a theorem.

THEOREM. *For* $0 \leqq Q \leqq 1$, *there exists a* β *for which the two-point boundary value problem* (1.1) *has at least one solution* f. *Further, on* $(0, 1)$: $-\frac{1}{3} \leqq f(\eta) < 0$, $-\frac{2}{3} \leqq f'(\eta) \leqq 1$, *and* $-\frac{2}{3} \leqq f''(\eta) \leqq 2$.

Remark. We observe that the β for which the above-described solution exists is given by

$$\beta = 2g'(0) - Qf''(0)^2 = g'(1) + 1 - Qf'(1)^2.$$

We can easily obtain crude bounds for β from Lemma 7 and the bounds on f'' given in the theorem: for $0 \leqq Q < 1.0$, $-Qf''(0)^2 \leqq \beta \leqq 1 + g'(1)$, or $2(1 + \frac{1}{3}Q) \geqq \beta \geqq -4Q$. For $Q = 0$, the upper bound is sharp: if $Q = 0$, $\beta = 2$. But as Q is increased from zero,

the upper bound becomes progressively worse when compared to the numerical results (see (Fig. 1). The upper bound 1.0 on Q in our theorem falls far short of 32.7, the actual numerically found upper bound for existence of solutions having the properties indicated in the theorem. Recently, R. Seydel of the University of Würzburg, Federal Republic of Germany, has informed the second author that he has found a third branch of solutions of (1.1), connecting the two branches in Fig. 1 and running between points not too far from their nodes.

REFERENCES

[1] W. N. GILL, N. D. KAZARINOFF, C. HSU, M. NOACK, AND J. D. VERHOEVEN, *Thermo-capillary driven convection in supported and floating zone crystallization*, Adv. in Space Res., 4 (1984), pp. 15–22.

[2] C. LU, N. D. KAZARINOFF, J. B. MCLEOD, AND W. C. TROY, *Existence of solutions of the similarity equations for floating rectangular cavities and disks*, SIAM J. Math. Anal., 19 (1988), pp. 1119–1126.

[3] J. SMOLLER, *Shock Waves and Reaction Diffusion Equations*, Grundlehren der Math. Wissenschaften 258, Springer-Verlag, Berlin, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 20, No. 2, pp. 504-510, March 1989

017

CONSTANT REGRESSION POLYNOMIALS AND THE WISHART DISTRIBUTION*

DONALD ST. P. RICHARDS†

Abstract. Results are obtained for the problems of constructing and characterizing scalar-valued polynomial statistics having constant regression on the mean of a random sample of Wishart matrices. The construction procedure introduced by Heller [*J. Multivariate Anal.*, 14 (1984), pp. 101-104] is generalized to show that certain polynomials in the principal minors of the sample matrices have zero regression on the mean. The zero-regression polynomials are characterized through expectations involving certain matrix-valued Bessel functions of Gross and Kunze [*J. Funct. Anal.*, 22 (1976), pp. 73-105]. It is shown that the zero-regression property characterizes Wishart distributions within a wide family of mixtures of Wishart distributions.

Key words. Wishart distribution, constant regression, polynomial statistics, hyperbolic differential operator, unitary representation, generalized Bessel function

AMS(MOS) subject classifications. primary 62H05, 62H10; secondary 33A65, 43A80

1. Introduction. Suppose that $X_1, X_2, \cdots, X_n$ are independent, identically distributed $m \times m$ random matrices, each having the Wishart distribution $W_m(\Sigma, N)$ with positive definite (symmetric) covariance matrix Σ and N degrees of freedom. A scalar-valued polynomial statistic $P(\underset{\sim}{X})$, $\underset{\sim}{X} = (X_1, \cdots, X_n)$, is said to have *constant regression* on $L = X_1 + \cdots + X_n$ if the conditional expectation

$$E(P(\underset{\sim}{X}) \mid L) = \beta \tag{1.1}$$

almost everywhere, for some constant β. Without loss of generality, we may suppose $\beta = 0$, in which case the polynomial $P(\underset{\sim}{X})$ is called a *zero-regression polynomial.* In this paper, we consider the problems of constructing and characterizing zero-regression polynomials.

In the classical case, Lukacs and Laha [8] have obtained necessary and sufficient conditions for a quadratic polynomial statistic to have zero regression. Their results have been presented, within the more general context of characterizations of probability distributions, by Kagan, Linnik, and Rao [7]. Although these problems were originally motivated by a purely academic desire to determine the theoretical distribution of the parent population from hypothetical properties of a particular statistic, recent applications [9] include the construction of testing procedures.

Recently, Heller [4] has constructed zero-regression polynomials for the Wishart distribution by appropriately generalizing the methods of [8]. Her results are among a small number that have extended the classical developments to the setting of matrix distributions.

The main tool used in [4] is the hyperbolic differential operator of Herz [5]. Using more complex operators, we generalize (in § 2) the construction procedure of [4]. As a consequence, we find that certain polynomials in the principal minors of the sample matrices are zero-regression polynomials.

Section 3 presents several characterizations of zero-regression polynomials using unconditional expectations. The main result characterizes zero-regression polynomials through expectations involving the matrix-valued Bessel functions of Gross and Kunze

* Received by the editors July 15, 1983; accepted for publication (in revised form) May 4, 1988.

† Department of Mathematics, University of Virginia, Charlottesville, Virginia 22903. The research of this author was supported by National Science Foundation grant MCS-8403381.

[2]. This result presents a new and unexpected link between harmonic analysis and multivariate statistical theory; further, it demonstrates the full complexity of the zero-regression problem in higher dimensions.

Following the results of [4], [8], a natural question is: Given that a polynomial $P(\underset{\sim}{X})$ has zero regression on L, is the underlying distribution necessarily Wishart? We obtain a partial answer, showing that if the parent distribution belongs to a wide class of Wishart mixtures, then the distribution is necessarily Wishart if $P(\underset{\sim}{X})$ is any one of a family of polynomials. In particular, this class of polynomials contains all the examples constructed in [4].

As a final comment, we remark that since the homogeneous zero regression polynomials with fixed degree form a finite-dimensional vector space, it would perhaps be more useful to determine the dimension of the space along with a "natural" basis. These problems have been solved by Kushner [14].

2. Construction of zero-regression polynomials. A partition $\kappa = (k_1, k_2, \cdots, k_m)$ is a vector of nonnegative integers, with $k_1 \geqq k_2 \geqq \cdots \geqq k_m$. For any complex number t,

$$(t)_\kappa = \prod_{i=1}^{m} \left(t - \frac{1}{2}(i-1) \right)_{k_i}$$

where

$$(t)_j = \frac{\Gamma(t+j)}{\Gamma(t)}, \qquad j = 0, 1, 2, \cdots.$$

Further, define the multivariate gamma function [5], [6] by

$$\Gamma_m(t) = \pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma\left(t - \frac{1}{2}(i-1) \right), \qquad \operatorname{Re}(t) > \tfrac{1}{2}(m-1).$$

For any symmetric $m \times m$ matrix $T = (t_{ij})$, we define

$$|T|_\kappa = \prod_{i=1}^{m} |T_i|^{k_i - k_{i+1}}, \qquad k_{m+1} \equiv 0,$$

where T_i is the ith principal minor of T, and $|\cdot|$ denotes the determinant. With the matrix T, we associate a matrix of differential operators

$$\frac{\partial}{\partial t} := \left(\frac{1}{2}(1+\delta_{ij}) \frac{\partial}{\partial t_{ij}} \right),$$

where δ_{ij} is Kronecker's delta. Whenever the variables of differentiation are clear from the context, we denote $\partial / \partial T$ by D; further, we shall use the notation

$$|D|_\kappa = \prod_{i=1}^{m} |D_i|^{k_i - k_{i+1}}$$

where D_i is the ith principal minor of D. For example, if $\kappa = (4, 2, 0, \cdots, 0)$, then

$$|D|_\kappa = \frac{\partial^2}{\partial t_{11}^2} \left(\frac{\partial^2}{\partial t_{11}\, \partial t_{22}} - \frac{\partial^2}{4 \partial t_{12}^2} \right)^2.$$

The operator $|D| \equiv |D|_{(1,1,\cdots,1)}$ is the hyperbolic operator of Herz [5].

If a random matrix X_1 has $W_m(\Sigma, N)$ distribution, then it is known that the characteristic function of X_1 is

$$\phi(T) := E(e^{i \operatorname{tr}(TX_1)}) = |I - 2iT\Sigma|^{-N/2},$$

where T is a symmetric matrix. Our first result determines the effect of the $|D|_\kappa$ operators on $\phi(T)$.

PROPOSITION 2.1. $|D|_\kappa \phi(T) = i^k (N/2)_\kappa |(\frac{1}{2}\Sigma^{-1} - iT)^{-1}|_\kappa \phi(T)$, *where* $k = k_1 + k_2 + \cdots + k_m$.

In the course of proving Proposition 2.1, we will use the following result by Bellman [1].

LEMMA 2.2 (Bellman). *Let Z be a complex, symmetric, $m \times m$ matrix with* $\operatorname{Re}(Z) > 0$ (*positive-definite symmetric*). *Then,*

$$\int_{X>0} e^{-\operatorname{tr}(XZ)} |X|^{t-p} |X|_\kappa \, dX = (t)_\kappa \Gamma_m(t) |Z|^{-t} |Z^{-1}|_\kappa,$$

with absolute convergence for $\operatorname{Re}(t) > p-1$, $p = (m+1)/2$.

Proof of Proposition 2.1. Let

$$C = (|2\Sigma|^{N/2} \Gamma_m(N/2))^{-1}$$

be the normalizing constant for the Wishart density. Then,

$$|D|_\kappa \phi(T) = C |D|_\kappa \int_{X>0} e^{i \operatorname{tr}(TX)} \, e^{-1/2 \operatorname{tr}(\Sigma^{-1}X)} |X|^{(N-m-1)/2} \, dX.$$

Since

$$|D|_\kappa \, e^{i \operatorname{tr}(TX)} = i^k |X|_\kappa \, e^{i \operatorname{tr}(TX)},$$

then

$$|D|_\kappa \phi(T) = C i^k \int_{X>0} e^{-\operatorname{tr}(1/2\Sigma^{-1} - iT)X} |X|^{(N-m-1)/2} |X|_\kappa \, dX.$$

Applying Lemma 2.2 and simplifying completes the proof. □

If $\kappa = (k_1, \cdots, k_m)$ and $\lambda = (l_1, \cdots, l_m)$ are partitions, we define $\kappa + \lambda = (k_1 + l_1, \cdots, k_m + l_m)$.

THEOREM 2.3. *Let $\{a_{ij}\}_{i,j=1}^n$ be real numbers. For any two partitions κ, λ define*

$$b_{\kappa,\lambda} = (N/2)_\kappa (N/2)_\lambda / (N/2)_{\kappa+\lambda}.$$

If $X_1, \cdots, X_n$ is a random sample from $W_m(\Sigma, N)$, then the polynomial

$$P(\underset{\sim}{X}) = \sum_{i \neq j}^{n} \sum^{n} a_{ij} [|X_i|_\kappa \, |X_j|_\lambda - b_{\kappa,\lambda} |X_i|_{\kappa+\lambda}] \tag{1}$$

is a zero-regression polynomial.

Proof. A necessary and sufficient criterion [4] for a polynomial $P(\underset{\sim}{X})$ to have zero regression on L is that

$$E(e^{i \operatorname{tr}(TL)} P(\underset{\sim}{X})) = 0 \tag{2}$$

for all symmetric T. To apply (2) to the polynomial (1), we shall use the following results:

$$|D|_\kappa \phi(T) = i^k E(e^{i \operatorname{tr}(TX)} |X|_\kappa),$$

$$|D|_\kappa |D|_\lambda \phi(T) = i^{k+l} E(e^{i \operatorname{tr}(TX)} |X|_\kappa |X|_\lambda),$$

where $k = k_1 + \cdots + k_m$, $l = l_1 + \cdots + l_m$. Then, applying (2) to (1), we obtain

$$E(e^{i \operatorname{tr}(TL)} P(\underset{\sim}{X})) = i^{-k-l} \sum_{r \neq s} \sum a_{rs} (\phi(T))^{n-2} \cdot [(|D|_\kappa \phi(T))(|D|_\lambda \phi(T)) - b_{\kappa,\lambda} \phi(T) |D|_{\kappa+\lambda} \phi(T)].$$

However, from Proposition 2.1, we also have

$$(|D|_\kappa \phi)(|D|_\lambda \phi) - b_{\kappa,\lambda}\phi |D|_{\kappa+\lambda}\phi \equiv 0,$$

and hence $P(\underset{\sim}{X})$ has zero regression on L. □

The construction given in Theorem 3.2 leads to a generalization of the results in [4]. To recover the zero regression polynomials of [4] from our result, we need only set $\kappa = \lambda = (1^m)$; this shows that

$$P(\underset{\sim}{X}) = \sum\sum_{i \neq j} a_{ij}[|X_i||X_j| - b_1|X_i|^2] \tag{3}$$

has zero regression on L, where

$$b_1 = (N/2)_{(1^m)}(N/2)_{(1^m)}/(N/2)_{(2^m)}.$$

3. Characterizations of zero-regression polynomials. This section characterizes zero-regression polynomials through several unconditional expectations, including generalizations of (2). First, we need to introduce notation and results pertinent to Fourier analysis on compact topological groups; an introductory account of this theory is given in [12] and [13].

Let G be a compact (Hausdorff) group. A *representation* ρ of G is a continuous homomorphism of G into the group of invertible linear transformations on a complex vector space V_ρ.

A representation ρ is *irreducible* if the only proper invariant (under ρ) subspace of V_ρ is $\{0\}$, the trivial subspace. Since G is compact, then V_ρ is necessarily of finite dimension, say d_ρ. Further, we lose no generality in assuming that the irreducible representations of G are given by unitary matrices.

Two representations ρ_1 and ρ_2 are *unitarily equivalent* if there exists a unitary matrix u such that $u\rho_1(g) = \rho_2(g)u$ for all g in G. The notion of unitary equivalence defines an equivalence relation among the representations of G. We denote by $\hat{G}$ the set of equivalence classes of *irreducible* representations of G. $\hat{G}$ is called the *dual* of G.

Let dg be the unique Haar measure on G, normalized so that G has total mass one. If $f: G \to \mathbb{C}$ is continuous, the *Fourier transform* of f is defined by

$$\rho(f) = \int_G f(g)\rho(g)\, dg, \qquad \rho \in \hat{G}.$$

The Fourier inversion formula then states that

$$f(g) = \sum_{\rho \in \hat{G}} d_\rho \operatorname{tr}(\rho(g)^* \rho(f)), \qquad g \in G, \tag{4}$$

where * denotes the transpose of complex conjugates. The sum in (4) is over a set of representatives for the equivalence classes in $\hat{G}$, one for each class.

Next, we specialize to the case $G = \mathrm{SO}(m)$, the *special orthogonal group*. The group G consists of all $m \times m$ orthogonal matrices g having determinant one.

Definition 3.1 [2], [3]. For symmetric $m \times m$ matrices S, T,

$$J_\rho(S, T) = \int_{\mathrm{SO}(m)} e^{i \operatorname{tr}(g'SgT)} \rho(g)\, dg$$

is the *generalized Bessel function of order* ρ.

The generalized Bessel functions were introduced and studied in [2]. In the jargon of [2], [3], $J_\rho(S, T)$ is the generalized Bessel function of order ρ arising from the two-sided action of $\mathrm{SO}(m)$ on the space of symmetric matrices.

Note that $J_\rho(S, T)$ is the Fourier transform of the function $g \to e^{i \operatorname{tr}(g'SgT)}$, $g \in \mathrm{SO}(m)$. Further, $J_\rho(S, T)$ is a $d_\rho \times d_\rho$ matrix. The Fourier inversion formula (4) shows that for fixed S, T,

$$e^{i \operatorname{tr}(g'SgT)} = \sum_{\rho \in \hat{G}} d_\rho \operatorname{tr}(\rho(g)^* J_\rho(S, T)). \tag{5}$$

Since the function $g \to e^{i \operatorname{tr}(g'SgT)}$ is infinitely differentiable, the series (5) converges uniformly on $\mathrm{SO}(m)$.

Now, we can state the main result of the paper. As before, $L = X_1 + \cdots + X_n$, where the X_i are independently and identically distributed Wishart matrices. Also, the polynomial $P(\underset{\sim}{X})$, $\underset{\sim}{X} = (X_1, \cdots, X_n)$, has zero regression on L.

THEOREM 3.2. *The following conditions are equivalent*:

(a) $P(\underset{\sim}{X})$ *has zero regression on* L;

(b) $E(e^{i \operatorname{tr}(TL)} p(L) P(\underset{\sim}{X})) = 0$ *for all polynomials* $p(L)$ *and symmetric matrices* T;

(c) $E(p(L) P(\underset{\sim}{X})) = 0$ *for all polynomials* $p(L)$;

(d) $E(P(\underset{\sim}{X}) J_\rho(T, L)) = 0$ *for all* $\rho \in \widehat{(\mathrm{SO}(m))}$ *and symmetric* T.

Proof. In § 2, we noted that the zero-regression property is equivalent to

$$E(e^{i \operatorname{tr}(TL)} P(\underset{\sim}{X})) = 0 \tag{6}$$

for all T. Then, it is immediate that (a) implies (b) since

$$E(e^{i \operatorname{tr}(TL)} p(L) P(\underset{\sim}{X})) = p\left(\frac{1}{i} \frac{\partial}{\partial T}\right) E(e^{i \operatorname{tr}(TL)} P(\underset{\sim}{X})).$$

Further, (b) implies (6) trivially, so that (a) and (b) are equivalent. To see that (c) implies (6), note that $e^{i \operatorname{tr}(TL)}$ can be expressed as an absolutely convergent series,

$$e^{i \operatorname{tr}(TL)} = \sum_{\alpha \in A} q_\alpha(T) p_\alpha(L)$$

where A is a countable index set, and p_α and q_α are polynomials for all α. Taking expectations, we find that for T sufficiently close to the zero matrix,

$$E(e^{i \operatorname{tr}(TL)} P(\underset{\sim}{X})) = \sum_{\alpha \in A} q_\alpha(T) E(p_\alpha(T) P(\underset{\sim}{X})) = 0$$

where the interchange of summation and expectation is permitted by the Dominated Convergence Theorem. As a function of T, the left-hand side of (6) is analytic and is zero in a neighborhood of the origin; hence it is identically zero. Since (b) implies (c), then we have established the equivalence of (a), (b), and (c).

Finally, we prove that (6) is equivalent to (d). First, Fubini's theorem and the definition of $J_\rho(S, T)$ show that

$$E(P(\underset{\sim}{X}) J_\rho(S, T)) = \int_{\mathrm{SO}(m)} E(e^{i \operatorname{tr}(g'TgL)} P(\underset{\sim}{X})) \rho(g)\, dg.$$

Therefore, (6) implies (d). Conversely if (d) holds, then by the Fourier inversion formula (5) and the Dominated Convergence Theorem,

$$E(e^{i \operatorname{tr}(g'TgL)} P(\underset{\sim}{X})) = \sum_\rho d_\rho \operatorname{tr}(\rho(g)^* E(P(\underset{\sim}{X}) J_\rho(T, L))) = 0,$$

for any g in $\mathrm{SO}(m)$ and symmetric T. Hence (d) implies (6), and this completes the proof. □

On reviewing the proof of Theorem 3.2, we note that nowhere do we use the fact that the X_i are Wishart matrices. Therefore the result holds for *any* symmetric matrix distribution, as long as the various interchanges of sums and integrals remain valid. In particular, it holds for the multivariate beta and F distributions [6].

To end this section, we make some remarks on the generalized Bessel functions. The *spherical Bessel function* is

$$J_0(S, T) = \int_{SO(m)} e^{i \operatorname{tr}(g'SgT)}\, dg,$$

which corresponds to the trivial one-dimensional representation $\rho(g) \equiv 1$. When $m = 2$, it can be shown using the techniques of [3] that

$$J_0(S, T) = e^{i \operatorname{tr}(S) \operatorname{tr}(T)/2} \tilde{J}_0(\Delta(S)\Delta(T)/2)$$

where $\tilde{J}_0(\cdot)$ denotes the classical Bessel function of the first kind of order zero, and $\Delta(S) = |s_1 - s_2|$ where s_1 and s_2 are the two eigenvalues of S. For any odd $m \geqq 3$, it has been shown [11] that $J_0(S, T) = {}_0F_0(iS, T)$, where ${}_0F_0(\cdot, \cdot)$ is James' hypergeometric function of two matrix arguments. For even $m \geqq 4$, $J_0(\cdot, \cdot)$ can be again related to ${}_0F_0(\cdot, \cdot)$ but the result is more complicated. For general ρ the evaluation of the matrix elements of $J_\rho(S, T)$ requires powerful machinery from the theory of harmonic analysis [2], [3]; however, we also remark that for the case $m = 2$, $J_\rho(S, T)$ can be computed explicitly using the techniques of [3].

4. A characterization of the Wishart distribution. Both here and in [4], it has been shown that the polynomial (3) has zero regression on $L = X_1 + \cdots + X_n$. Given the converse, that (3) has zero regression on the sum of an independently and identically distributed sequence of random matrices, we wish to determine whether the underlying population is necessarily Wishart. From the proof of Theorem 2.3, we observe that the polynomial (3) has zero regression on L if and only if the characteristic function $\phi(T) = E(e^{i \operatorname{tr}(TX_1)})$ satisfies the differential equation

$$(|D|\phi(T))^2 = b_1 \phi(T) |D|^2 \phi(T), \tag{7}$$

for all T. Therefore, it is enough to find all solutions of (7) that are simultaneously characteristic functions.

Consider the class of characteristic functions $\phi(T)$ of the form

$$\phi(T) = \int_H |I - 2iT\Sigma|^{-N/2}\, d\nu(\Sigma) \tag{8}$$

where $\nu(\cdot)$ is a Borel probability measure on the hypersurface $H = \{\Sigma: \Sigma \text{ is positive definite and } |\Sigma| = 1\}$. A random matrix with a characteristic function of form (8) may be regarded as a "covariance mixture" of Wishart distributions; these mixtures have arisen [10] in the context of hyperspherical distributions. We now show that the zero regression polynomial (3) characterizes the Wishart distribution within the class of Wishart mixtures typified by (8).

THEOREM 4.1. *Let $X_1, \cdots, X_n$ be a random sample from a population with characteristic function* (8). *If the polynomial $P(\underset{\sim}{X})$ in* (3) *has zero regression on L, then X_1 has a Wishart distribution.*

Proof. Since $P(\underset{\sim}{X})$ has zero regression on L, then (7) holds. Applying the $|D|$ and $|D|^2$ operators to $\phi(T)$ in (8) shows that

$$\left(\int_H |I - 2iT\Sigma|^{-(N+2)/2}\, d\nu(\Sigma)\right)^2 = \int_H |I - 2iT\Sigma|^{-N/2}\, d\nu(\Sigma) \cdot \int_H |I - 2iT\Sigma|^{-(N+4)/2}\, d\nu(\Sigma), \tag{9}$$

for all symmetric T. By considering discrete approximations to these latter integrals, it is evident that (9) implies $\nu(\cdot)$ to be the Dirac measure at some "point" Σ_0. Consequently, $X_1 \sim W_m(N, \Sigma_0)$.

Stephen Marron has kindly remarked that the above conclusion concerning $\nu(\cdot)$ can also be obtained by working with moment generating functions rather than characteristic functions. Indeed, the generating function analogue of (9) is

$$(9') \qquad \left(\int_H |I-2T\Sigma|^{-(N+2)/2}\, d\nu(\Sigma)\right)^2 = \int_H |I-2T\Sigma|^{-N/2}\, d\nu(\Sigma) \cdot \int_H |I-2T\Sigma|^{-(N+4)/2}\, d\nu(\Sigma),$$

and this holds for all T in a sufficiently small neighborhood of the zero matrix. The Cauchy-Schwarz inequality guarantees that the left-hand side of (9′) is never larger than the right-hand side. Then, the conclusion that $\nu(\cdot)$ is Dirac follows immediately from the criterion for equality in the Cauchy-Schwarz inequality. □

Finally, we remark that results similar to Theorem 4.1 can be obtained for the general polynomials constructed in Theorem 2.3.

Acknowledgments. I am deeply grateful to Kenneth Gross for his patient explanations of harmonic analysis in general, and the generalized Bessel functions in particular.

REFERENCES

[1] R. E. BELLMAN, *A generalization of some integral identities due to Ingham and Siegel*, Duke Math. J., 23 (1956), pp. 571–577.

[2] K. I. GROSS AND R. A. KUNZE, *Bessel functions and representation theory*, J. Funct. Anal., 22 (1976), pp. 73–105.

[3] K. I. GROSS, W. J. HOLMAN, III, AND R. A. KUNZE, *A new class of Bessel functions and applications in harmonic analysis*, in Proc. Symposia in Pure Mathematics, XXXV, American Mathematical Society, Providence, RI, 1979, pp. 407–415.

[4] B. HELLER, *Use of the hyperbolic operator to find a scalar statistic which has constant regression on the mean of a sample of Wishart matrices*, J. Multivariate Anal., 14 (1984), pp. 101–104.

[5] C. S. HERZ, *Bessel functions of matrix argument*, Ann. of Math., 61 (1955), pp. 474–523.

[6] A. T. JAMES, *Distributions of matrix variates and latent roots derived from normal samples*, Ann. Math. Statist., 35 (1964), pp. 475–501.

[7] A. M. KAGAN, Y. V. LINNIK, AND C. R. RAO, *Characterization Problems in Mathematical Statistics*, John Wiley, New York, 1973.

[8] E. LUKACS AND R. G. LAHA, *Applications of Characteristic Functions*, Griffin, London, 1964.

[9] G. S. MUDHOLKAR AND C. C. LIN, *A simple test for normality against asymmetric alternatives*, Biometrika, 67 (1980), pp. 455–461.

[10] D. ST. P. RICHARDS, *Hyperspherical models, fractional derivatives and exponential distributions on matrix spaces*, Sankhyā. Ser. A, 46 (1984), pp. 155–165.

[11] ———, *Solution to Problem* 84-1: *An integral on* SO(3), SIAM Rev. 27 (1985), pp. 81–82.

[12] G. WEISS, *Harmonic analysis on compact groups*, in Studies in Harmonic Analysis, J. M. Ash, ed., Mathematical Association of America, Providence, RI, 1976, pp. 198–223.

[13] M. A. NAIMARK AND A. I. STERN, *Theory of Group Representations*, Springer-Verlag, Berlin, New York, 1982.

[14] H. B. KUSHNER, *Dimensions of spaces of homogeneous zero regression polynomials*, J. Multivariate Anal., 22 (1987), pp. 245–250.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 511-532, May 1989

HOPF BIFURCATION IN THE PRESENCE OF SPHERICAL SYMMETRY: ANALYTICAL RESULTS*

G. IOOSS†‡ AND M. ROSSI‡

Abstract. This paper considers a one-parameter family of vector fields equivariant under the orthogonal group $O(3)$, with an invariant fixed point. Hopf bifurcation of this family is studied assuming $O(3)$ acts on the eigenspaces belonging to each purely imaginary eigenvalue as an $l=2$ representation. The dynamics are then reduced to a ten-dimensional center manifold, and the normal form of the vector field is explicitly given up to fifth order. The five different types of bifurcating periodic solutions, predicted geometrically by Golubitsky and Stewart, are derived analytically: a family of axisymmetric solutions, two types of rotating waves (one rotating at twice the speed of the other), a family of standing waves and a family of tetrahedral waves.

The stability conditions are given for all these solutions. These stabilities depend on the three coefficients appearing at cubic order in the normal form and on one combination of three coefficients occurring at fifth order. All solutions can be stable except for the fastest family of rotating waves. The slower rotating waves and the axisymmetric solutions may be simultaneously stable. Finally it is shown that a family of quasiperiodic solutions may bifurcate directly from the invariant fixed point together with the periodic solutions.

Key words. Hopf bifurcation symmetry breaking, spherical symmetry

AMS(MOS) subject classifications. 58F, 58C

1. Introduction.

Systems possessing a priori a spherical symmetry frequently occur in physics, particularly in hydrodynamics (see, for instance, self-gravitation convection problems [3], and the evolution of the shape of a gas bubble in a liquid [16]). Such phenomena obey a system of partial differential equations of evolution type, which may be written in the form of a differential equation:

$$\frac{dU}{dt} = \mathcal{F}(\lambda, U) \tag{1}$$

in a suitable functional real space E (see [10] for a precise formulation in hydrodynamics. Here $U(t)$ may stand for various physical quantities at time t: the velocity vector field of fluid particles, temperature, location of a free surface, etc. Moreover, $\lambda\ \mathbb{R}^k$ characterizes all the parameters of the underlying physics. In this mathematical frame, spherical symmetry means that $\mathcal{F}(\lambda, \cdot)$ commutes with a representation of the orthogonal group $O(3)$, i.e.,

$$\mathcal{F}(\lambda, \gamma U) = \gamma \mathcal{F}(\lambda, U) \tag{2}$$

for any element γ of the representation of $O(3)$ on the space E. Let us assume that we know a steady solution of (1) invariant under $O(3)$. We take it to be the origin in E. So, we have:

$$\mathcal{F}(\lambda, 0) = 0. \tag{3}$$

We assume in the following that this solution is marginally stable when λ equals zero. More precisely, the spectrum of the linear operator $\mathcal{L}_\lambda = D_U \mathcal{F}(\lambda, \cdot)$ is divided into two parts: one part, denoted by Σ_0, lying on the imaginary axis; the other lying strictly to the left side of this axis. In a neighborhood of $\lambda = 0$, the center manifold theorem implies that the dynamics of (1) near zero in space E is asymptotically

* Received by the editors October 14, 1987; accepted for publication (in revised form) June 9, 1988.

† Laboratoire de Mathématiques, U.A. Centre National de Recherche Scientifique 168, Université de Nice, Parc Valrose 06034 Nice, France.

‡ Centre d'Etudes de Limeil-Valenton, B.P. 27, 94190 Villeneuve Saint Georges, France.

described by the trace of the vector field (1) on a center manifold $\mathcal{V}_0$. This manifold $\mathcal{V}_0$ has the same dimension as E_0 and is both locally invariant and locally attracting in a neighborhood of $U=0$. A complete modern proof of this theorem may be found in [19] for vector fields in finite-dimensional spaces. For evolution problems described by partial differential equations, see [11] or an easy adaptation of [19], which may be done by using, for instance, results reviewed in § II of [10]. This theorem is proved in an easier way for maps (see, for instance, [14], [9]), but then an adaptation for vector fields is needed (see also § V.4 of [9] for partial differential equations). The manifold $\mathcal{V}_0$ may be expressed in the following form:

$$U = X + \Phi(\lambda, X), \quad \Phi(0,0)=0, \quad D_X\Phi(0,0)=0, \tag{4}$$

where X belongs to the subspace E_0 invariant under $\mathcal{L}_0$ and corresponds to the part Σ_0 of the spectrum. In the problems considered, especially in hydrodynamics, this subspace is generally finite-dimensional. Therefore, after projection on the manifold, we are concerned only with a finite-dimensional differential system:

$$\frac{dX}{dt} = F(\lambda, X), \tag{5}$$

where $X(t) \in E_0$, $F(\lambda, 0)=0$.

The linear operator $L_0 = D_X F(0,0)$ is the restriction of $\mathcal{L}_0$ to E_0, so its entire spectrum is located on the imaginary axis. Moreover, the equivariance (2) of $\mathcal{F}$ remains valid for the new vector field on E_0, since a theorem of Ruelle [18] shows that we can find Φ such that

$$F(\lambda, \gamma X) = \gamma F(\lambda, X), \tag{6}$$

$$\Phi(\lambda, \gamma X) = \gamma \Phi(\lambda, X), \tag{7}$$

where γ is an element of the representation Γ of $O(3)$ on E_0. Straightforward consequences of (6) and (7) are: (a) L_0 commutes with every element of the representation Γ; (b) the eigenvalues of L_0 may have a large multiplicity; and (c) E_0 may be large-dimensional. Hence even for simple generic bifurcations of codimension 1, numerous solutions may bifurcate with complex patterns.

Using geometrical arguments from Lie group theory, Golubitsky and Stewart [7] have considered the "simplest" case of a Hopf bifurcation in the presence of $O(3)$ symmetry: they assumed that L_0 possesses a pair of eigenvalues $\pm i\omega$ associated with the decomposition of the complexified space E_0^c denoted hereafter by E_0:

$$E_0 = V \oplus \bar{V},$$

where the representation of $O(3)$ on V (and $\bar{V}$) is absolutely irreducible. Because of this assumption, the above-mentioned authors showed the emergence of branches of bifurcated solutions in each two-dimensional space corresponding to some specified symmetry. The method of restricting the study to subspaces of minimal dimensionality was previously used by Iudovich [12] to search analytically for steady solutions in Bénard convection. The geometric arguments developed in [7] do not say anything about the existence of other solutions (periodic or quasiperiodic), and do not show explicitly the stability of the specific periodic solutions.

Our purpose is to obtain more insight into these questions by studying the amplitude equations for a specific case. The case we investigate is mathematically the simplest leading to multiple solutions—among them one possessing an interesting tetrahedral symmetry. We note that nonstationary solutions might also be found in

cases of mode interaction when the only eigenvalue of L_0 is zero, its multiplicity being high enough: 5 in [3], 8 in [20].

The paper is organized as follows. In § 1 we derive the form of the differential system (5) on the center manifold $\mathcal{V}_0$. Unlike the steady bifurcation case [8], the irreducible representation $l=2$ of $O(3)$ does not possess a simple expression for the Hopf problem on the subspace E_0. This, as well as the dimension $2(2l+1)=10$ of the differential system (5), requires us to make use of rather technical results in this section. Consequently, we have only summarized the essential features of the calculus to get the form (24). The following section reviews and describes explicitly the five time-periodic solutions predicted in [7]: one axisymmetric solution, two rotating waves solutions (one type rotating twice as rapidly as the other), one standing wave solution, and one tetrahedral wave solution. We look exhaustively for their stability and show in particular that the "fastest" rotating wave solution is always unstable and that several solutions may be simultaneously stable. All the stability conditions may be expressed with three coefficients from the third order and *only one* combination of coefficients from the fifth order. In the last section we investigate the possible existence of quasi-periodic solutions directly emerging after a codimension 1 bifurcation, together with periodic solutions.

2. Normal form of the amplitude equations.

2.1. The basic problem. Our purpose is to find, when λ is near zero, how to describe the asymptotic dynamics of the solutions of (5) whenever E_0 can be written in the form

$$E_0 = V \oplus \bar{V},$$

where V, of dimension five, possesses an irreducible representation $l=2$ for the group $O(3)$. Recall that F commutes with this representation.

Each real vector of E_0 may be decomposed as

$$X = \sum_{m=-2}^{m=+2} x_m \xi_m + \bar{x}_m \bar{\xi}_m, \tag{8}$$

where the ξ_m ($m=-2,-1,0,1,2$) are eigenvectors of L_0 for the eigenvalue $i\omega$. This decomposition differs slightly from that of [7]: Golubitsky and Stewart write their hypothesis on a real basis corresponding to the real and imaginary parts of our vectors. It follows from this that the problem now reduces to that of finding vector fields depending on complex amplitudes x_j, $\bar{x}_j$ ($j=-2,-1,0,1,2$) that satisfy (6), i.e., after differentiation, the following relations:

$$D_X F(\lambda, X) \cdot J_k X = J_k F(\lambda, X), \qquad k=1,2,3,$$

or

$$\begin{aligned} &D_X F(\lambda, X) \cdot J_+ X = F(\lambda, X), \\ &D_X F(\lambda, X) \cdot J_- X = J_- F(\lambda, X), \\ &D_X F(\lambda, X) \cdot J_3 X = J_3 F(\lambda, X), \end{aligned} \tag{9}$$

where $-iJ_1, -iJ_2, -iJ_3$ are infinitesimal generators corresponding to rotations about each coordinate axis, and

$$J_\pm = J_1 \pm iJ_2.$$

Remark. Only two of the three relations in (9) are independent, since if F commutes with rotations about two coordinate axes it commutes with the rotation about the third axis.

To derive the general amplitude equation consistent with (9), we must consider the fundamental relations of the irreducible $O(3)$ representation $l=2$. In the following paragraph, we briefly review these properties without proving them. We refer the reader to [6] or [15] for details.

2.2. Irreducible representation $l=2$ of the orthogonal group. The linear operators J_3, J_+, J_- act on the canonical basis $\{\xi_m, m=-2,-1,0,1,2\}$ as follows (see pp. 24-25 of [6]):

$$J_3\xi_m = m\xi_m, \quad J_-\xi_m = \beta_{-m}\xi_{m-1}, \quad J_+\xi_m = \beta_m\xi_{m+1} \tag{10}$$

where $\beta_m = \sqrt{(2-m)(3+m)}$ and $m \in \{-2,-1,0,1,2\}$. Similarly (since $\bar{J}_+ = -J_-$ and $\bar{J}_3 = -J_3$):

$$J_3\bar{\xi}_m = -m\bar{\xi}_m, \quad J_+\bar{\xi}_m = -\beta_{-m}\bar{\xi}_{m-1}, \quad J_-\bar{\xi}_m = -\beta_m\bar{\xi}_{m+1}. \tag{11}$$

In Appendix 1 we describe the action of any finite rotation in this representation. Let us point out here the special case of a rotation of angle π about the axis ox:

$$R_{ox}(\pi)\xi_m = \xi_{-m}, \tag{12}$$

and that of a rotation of angle θ about oz:

$$R_{oz}(\theta)\xi_m = e^{-im\theta}\xi_m. \tag{13}$$

The reflection S through the origin possesses two possible representations: Id and $-\mathrm{Id}$. The choice between these two possibilities depends on the particular physical problem. The natural representation on spherical harmonics $(-\mathrm{Id})^l$ corresponds to the identity when $l=2$. It turns out, however, that the other possibility just modifies the symmetry of the solutions but not the dynamical system itself. In § 2.3 we suppose the natural representation holds: straightforward modifications belonging to the other case are left as an exercise for the reader.

2.3. Normal form of the vector field on the center manifold. To analyze the amplitude equations, we must put them into normal form. The basic result, obtained by Elphick et al. [5], indicates the existence of a nonlinear change of variables (close to the identity) such that the normal form of F commutes with the one-parameter group $\exp(L_0 t)$. After differentiation this property may be put into a more convenient form:

$$D_X F(\lambda, X)\cdot L_0 X = L_0 F(\lambda, X), \tag{14}$$

where we recall that the action of L_0 on the canonical basis is

$$L_0 = i\omega\,\mathrm{Id}_V. \tag{15}$$

Relation (14) is, however, only applicable up to an arbitrarily large but fixed order. Indeed, it is known that normalization of a vector field cannot be done to all orders, due to problems of convergence. The degree of the polynomial F must first be fixed (large N); then the neighborhood of zero, where an estimate $O(\|X\|^N)$ on the higher-order terms still holds, is fixed. This is not the case for the properties (9) of F that are valid exactly, due to the fundamental equivariance (2) of the system and to the results of [18]. From now on, we will say that a property is verified "up to flat terms" if it is verified on polynomial F.

2.4. Computation of the normal form. General approach. To determine the most general expression for F compatible with (9) and (14) let us decompose $F(\lambda, X)$. This yields:

$$F(\lambda, X) = \sum_{m=-2}^{m=+2} F_m\xi_m + \bar{F}_m\bar{\xi}_m, \tag{16}$$

where each component is explicitly expanded in the power series

$$F_m = \sum_P \alpha_m^P x_{-2}^{p_{-2}} \cdots x_2^{p_2} \bar{x}_{-2}^{q_{-2}} \cdots \bar{x}_2^{q_2} \tag{17}$$

with $P = (p_{-2}, \cdots, p_2, q_{-2}, \cdots, q_2)$.

The equalities $(9)_3$ and (14) then select those α_m^P that verify the resonant conditions:

$$\sum_{j=-2}^{2} j(p_j - q_j) = m, \tag{18}$$

$$\sum_{j=-2}^{2} (p_j - q_j) = 1. \tag{19}$$

Remark 1. Since we are looking for a polynomial F, we do not distinguish between the type of conditions generated by $(9)_3$ and (14).

Remark 2. The consequences of $(9)_1$ and $(9)_2$ do not give such simple relations since J_+, J_- are not diagonal operators on the canonical basis.

The general form of F is computed in two steps. First we determine F_{-2}. In addition to (18) and (19), $(9)_1$ implies that the normal form must satisfy

$$D_X F_{-2} \cdot J_+ X = 0. \tag{20}$$

Assume this calculation is done: then we easily obtain F_{-1} and F_0 thanks to $(9)_2$:

$$\beta_1 F_{-1} = D_X F_{-2} \cdot J_- X, \tag{21}$$

$$\beta_0 F_0 = D_X F_{-1} \cdot J_- X. \tag{22}$$

Now, instead of using J_- again to obtain F_1, F_2, we use a shortcut based on the equivariance under the rotation $R_{ox}(\pi)$. This gives F_1 and F_2, since (12) and (16) lead to

$$F_m(\lambda, R_{ox}(\pi)X) = F_{-m}(\lambda, X). \tag{23}$$

Other relations from (9), (18), (19) are then automatically satisfied and the most general normal form is thus found.

2.5. Computation of F_{-2}. To exhibit F_{-2} it is possible to use the pedestrian way of examining by increasing degrees the monomials (17) in (20) and eliminating those incompatible with (9), (18), and (19). This method—although conceptually simple—can, however, hardly ever be applied beyond the third order because of the exponential growth of the number of monomials involved. (For the third-order there are 5,000 for only three vectorial resonating terms.) Furthermore it does not yield a general result on the form of F_{-2}. This difficult question has been solved using Lie theoretic techniques: Cerezo [1] shows, for instance, that the polynomial function F_{-2} may be written as a linear combination of 43 polynomials, the coefficients of which are invariant polynomials of the group $O(3) \times SO(2)$ (generators J_3, J_+, L_0). At order 7, 23 of the 43 terms are present in F_{-2}, and it is necessary to go to the thirteenth order to find all the possible terms. It would be rather tedious to reproduce these results (the interested reader should consult [1] where a thorough study of this matter is given). Here we give only the expression truncated to the fifth order: we must go to this order to analyze the degeneracy occurring in the study of stability of some of the bifurcated solutions.

In what follows we denote by $i\omega + \mu(\lambda)$ the eigenvalue of the operator $\mathscr{L}_\lambda$, which perturbs $i\omega$ when λ is different from zero. All the coefficients in the vector field $F(\lambda, X)$ are functions of λ, but just the dependency in λ of the linear term is explicitly written.

As a matter of fact, the nonlinear terms are supposed to be nonsingular for $\lambda = 0$. This means that only codimension 1 bifurcations are considered and justifies the replacement of λ by μ_r (bifurcation parameter). The complete calculation yields the following for the vector field F:

$$\begin{aligned} F(\lambda, X) = {} & [i\omega + \mu + a|X|^2 + d_1|X|^4 + d_2|S_1(X)|^2] + (b + d_4|X|^2)S_1(X)\tilde{X} \\ & + (c + d_3|X|^2)C(X) + d_5 S_1(X)\tilde{C}(X) + d_6 S_3(X)B_1(X) \\ & + d_7 S_2(X)\tilde{B}_1(X) + d_8 \bar{S}_2(X)B_2(X) + d_9 M(X) + O(|X|^7), \end{aligned} \tag{24}$$

where we denote again by F and X the following vectors (to simplify notation):

$$F = (F_{-2}, F_{-1}, F_0, F_1, F_2), \qquad X = (x_{-2}, x_{-1}, x_0, x_1, x_2).$$

The mapping $Y \to \tilde{Y}$ is defined in $\mathbb{C}^5$ by

$$Y = (y_{-2}, y_{-1}, y_0, y_1, y_2) \to \tilde{Y} = (\bar{y}_2, -\bar{y}_1, \bar{y}_0, -\bar{y}_{-1}, \bar{y}_{-2}).$$

In equality (24), there appear six $O(3)$ scalar invariant quantities $|X|^2$, $S_1(X)$, $\bar{S}_1(X)$, $S_2(X)$, $\bar{S}_2(X)$, $S_3(X)$ as defined below (in the full normal form of F there also appear $\bar{S}_3(X)$ and another real invariant of degree 4 (see [1])):

$$\begin{aligned} |X|^2 &= \sum_{m=-2}^{m=+2} |x_m|^2, \\ S_1(X) &= x_{-1}x_1 - \frac{1}{2}x_0^2 - x_{-2}x_2, \\ S_2(X) &= x_2\left(\sqrt{\frac{3}{2}}\,\bar{x}_1^2 - 2\bar{x}_0\bar{x}_2\right) + x_{-2}\left(\sqrt{\frac{3}{2}}\,x_{-1}^2 - 2\bar{x}_{-2}\bar{x}_0\right) + x_1(\bar{x}_0\bar{x}_1 - \sqrt{6}\bar{x}_{-1}\bar{x}_2) \\ &\quad + x_{-1}(\bar{x}_{-1}\bar{x}_0 - \sqrt{6}\bar{x}_{-2}\bar{x}_1) + x_0(-\bar{x}_{-1}\bar{x}_1 - 2\bar{x}_{-2}\bar{x}_2 + \bar{x}_0^2), \\ S_3(X) &= x_0^3 - 3x_{-1}x_0x_1 + 3\sqrt{\frac{3}{2}}\,(x_{-1}^2x_2 + x_{-2}x_1^2) - 6x_{-2}x_0x_2. \end{aligned} \tag{25}$$

To avoid lengthy expressions, let us now introduce the operator $Y \mapsto \hat{Y}$ in $\mathbb{C}^5$ defined by

$$\hat{Y} = (y_2, y_1, y_0, y_{-1}, y_{-2}).$$

This operator represents $R_{ox}(\pi)$ (see (12)). We may now define:

$$C(X) = (C_{-2}, C_{-1}, C_0, C_1, C_2), \qquad B_j(X) = (B_{-2}^{(j)}, B_{-1}^{(j)}, B_0^{(j)}, B_1^{(j)}, B_2^{(j)})$$

with $C_m(\hat{X}) = C_{-m}(X)$, $B_m^{(j)}(\hat{X}) = B_{-m}^{(j)}(X)$, and

$$C_{-2}(X) = x_{-1}^2\bar{x}_0 + x_{-1}x_0\bar{x}_1 + \frac{1}{\sqrt{6}}\,x_0^2\bar{x}_2 - \sqrt{6}x_{-2}\left(\frac{2}{3}|x_0|^2 + |x_1|^2 + |x_2|^2\right),$$

$$C_{-1}(X) = \sqrt{\frac{2}{3}}\,x_0^2\bar{x}_1 + x_0x_1\bar{x}_2 + x_{-2}x_1\bar{x}_0 + 2x_{-2}x_0\bar{x}_{-1} - \sqrt{\frac{3}{2}}\,x_{-1}\left(|x_{-1}|^2 + \frac{1}{3}|x_0|^2 + |x_1|^2 + 2|x_2|^2\right),$$

$$C_0(x) = (x_1^2\bar{x}_2 + x_{-1}^2\bar{x}_{-2}) + (x_{-2}x_1\bar{x}_{-1} + x_{-1}x_2\bar{x}_1) + \sqrt{\frac{2}{3}}\, x_{-2}x_2\bar{x}_0$$
$$+ 2\sqrt{\frac{2}{3}}\, x_{-1}x_1\bar{x}_0 - \frac{1}{\sqrt{6}}\, x_0[4|x_{-2}|^2 + |x_{-1}|^2 + 3|x_0|^2 + |x_1|^2 + 4|x_2|^2],$$

$$B_0^{(1)}(x) = \sqrt{\frac{2}{3}}\,(\bar{x}_0^2 - \bar{x}_{-1}\bar{x}_1 - 2\bar{x}_{-2}\bar{x}_2),$$

$$B_{-1}^{(1)}(x) = 2\bar{x}_{-1}\bar{x}_2 - \sqrt{\frac{2}{3}}\,\bar{x}_0\bar{x}_1,$$

$$B_{-2}^{(1)}(X) = \bar{x}_1^2 - 2\sqrt{\frac{2}{3}}\,\bar{x}_0\bar{x}_2,$$

$$B_0^{(2)}(X) = |x_{-2}|^2 - \frac{1}{2}|x_{-1}|^2 - |x_0|^2 - \frac{1}{2}|x_1|^2 + |x|_2^2,$$

$$B_{-1}^{(2)}(X) = \sqrt{\frac{3}{2}}\,(x_1\bar{x}_2 - x_{-2}\bar{x}_{-1}) + \frac{1}{2}(x_0\bar{x}_1 - x_{-1}\bar{x}_0),$$

$$B_{-2}^{(2)}(X) = x_{-2}\bar{x}_0 + x_0\bar{x}_2 + \sqrt{\frac{3}{2}}\, x_{-1}\bar{x}_1.$$

Finally, we obtain

$$M(X) = (M_{-2}, M_{-1}, M_0, M_1, M_2), \qquad M_{-j}(X) = M_j(\hat{X}),$$

with

$$M_{-2}(X) = D_{-2}Q_{-2},$$
$$M_{-1}(X) = D_{-1}Q_{-2} + D_{-2}Q_{-1},$$
$$M_0(X) = D_0Q_{-2} + 2\sqrt{\frac{2}{3}}\, D_{-1}Q_{-1} + D_{-2}Q_0,$$

where

$$D_0(X) = -\frac{1}{\sqrt{6}}\, D_{-2}(\hat{X}), \qquad Q_0(X) = -\frac{1}{\sqrt{6}}\, Q_{-2}(\hat{X}),$$

and

$$Q_{-2}(X) = \bar{x}_1x_0 + x_{-1}\bar{x}_0 + \sqrt{\frac{2}{3}}\,(x_1\bar{x}_2 + x_{-2}\bar{x}_{-1}),$$

$$Q_{-1}(X) = \frac{1}{\sqrt{6}}\,(|x_1|^2 - |x_{-1}|^2) + \sqrt{\frac{2}{3}}\,(|x_2|^2 - |x_{-2}|^2),$$

$$D_2(X) = \bar{x}_{-1}\left(2x_{-2}x_0 - \sqrt{\frac{3}{2}}\, x_{-1}^2\right) + \bar{x}_0\left(3x_{-2}x_1 - \sqrt{\frac{3}{2}}\, x_{-1}x_0\right)$$
$$- \sqrt{\frac{3}{2}}\,\bar{x}_1(x_0^2 - x_{-1}x_1 - 2x_{-2}x_2) + \bar{x}_2(\sqrt{6}x_{-1}x_2 - x_0x_1),$$

$$D_{-1}(X) = \bar{x}_{-2}\left(\sqrt{\frac{3}{2}}\, x_{-1}^2 - 2x_{-2}x_0\right) + \bar{x}_{-1}\left(\frac{1}{2}\, x_{-1}x_0 - \sqrt{\frac{3}{2}}\, x_{-2}x_1\right)$$
$$+ \bar{x}_1\left(\sqrt{\frac{3}{2}}\, x_{-1}x_2 - \frac{1}{2}\, x_0x_1\right) + \bar{x}_2\left(2x_0x_2 - \sqrt{\frac{3}{2}}\, x_1^2\right).$$

It is worth noting that (a) only odd degrees appear in (24) because of condition (19), and (b) only three terms are needed to describe F up to the fourth order. Nine additional terms must be included ($d_1, d_2, d_3, \cdots, d_9$) to improve the expansion up to the fifth order. Moreover, in polynomial coefficients of $d_6, \cdots, d_9$ new types of terms $\{S_2(X), S_3(X), B_1(X), B_2(X), M(X)\}$ occur, which do not exist up to third order; hence it is not possible to guess them starting with the knowledge of the cubic terms.

2.6. Effective dimension of the system. The amplitude equation (5) actually depends on eight rather than ten variables. If we express the amplitudes in polar form:

$$x_j = r_j e^{i\psi_j}, \qquad -2 \leqq j \leqq 2, \tag{26}$$

then conditions (18), (19) imply that, in (5), the phases ψ_j appear only inside three linearly independent combinations:

$$\psi_1 + \psi_{-1} - 2\psi_0, \quad \psi_2 + \psi_{-2} - 2\psi_0, \quad 2\psi_{-1} - \psi_0 - \psi_{-2}. \tag{27}$$

An immediate consequence of (27) is the following remark. After a change of variables

$$x_j = y_j e^{i\omega_j t}, \tag{28}$$

where the parameters ω_j are given arbitrarily, the new system remains *autonomous* when these parameters satisfy

$$\omega_j = \beta + j\alpha, \qquad j = -2, -1, 0, 1, 2, \quad \text{where } \alpha, \beta \text{ are real.} \tag{29}$$

The "flat terms," however, do not satisfy this property since we use (19) to derive (29). If we search for a global property we have only to take into account (18). Instead of (29) this leads to

$$\omega_j = j\alpha, \qquad j = -2, -1, 0, 1, 2, \quad \text{where } \alpha \text{ is real,} \tag{30}$$

and then the system is again *completely autonomous.*

3. The five predicted bifurcated solutions. Symmetry and stability. In this section we explicitly derive the five solutions geometrically predicted in [7] and we study their stabilities. Those five branches of solutions for $l = 2$ share the following feature. The vector space remaining invariant under the action of the spatio-temporal symmetry group attached to each branch is two-dimensional. Hence we search for solutions reducing (5) to a system of the same dimension.

3.1. Axisymmetric solution. If $x_j = 0$ for $j = -2, -1, 1, 2$, then (18), (19) lead to

$$F_m = 0 \quad \text{for } m = -2m - 1, 1, 2.$$

In fact, the nonzero monomials correspond to terms such that

$$m = 0, \qquad p_0 - q_0 = 1.$$

Hence (5) reduces to (apart from "flat terms"):

$$\frac{dx_0}{dt} = x_0 \tilde{F}(|x_0|^2). \tag{31}$$

Truncation at fifth order $F(\lambda, X)$ in (24) gives

$$\frac{dx_0}{dt} = x_0(i\omega + \mu + \chi^{(1)}|x_0|^2 + \chi^{(2)}|x_0|^4) \tag{32}$$

with

$$\chi^{(1)} = a - \frac{b}{2} - c\sqrt{\frac{3}{2}},$$

$$\chi^{(2)} = d_1 + \frac{d_2}{4} + \frac{1}{2}\sqrt{\frac{3}{2}}\, d_5 - \sqrt{\frac{3}{2}}\, d_3 - \frac{d_4}{2} + \sqrt{\frac{2}{3}}\,(d_6 + d_7) - d_8.$$

Generically this leads to a standard Hopf bifurcation. The time-periodic solution reads as follows:

$$x_0 = r_0\, e^{i(\omega_0 t + \psi_0)}, \qquad x_j = 0 \quad (j \neq 0), \tag{33}$$

where r_0, ω_0 (index i (respectively, r) denotes imaginary (respectively, real part)) are given by

$$\mu_r + \chi_r^{(1)} r_0^2 + \chi_r^{(2)} r_0^4 + O(r_0^6) = 0, \qquad \omega_0 = \omega + \mu_i + \chi_i^{(1)} r_0^2 + \chi_i^{(2)} r_0^4 + O(r_0^6). \tag{34}$$

In space E, the bifurcated solution (33) takes the following form (up to an arbitrary phase corresponding to the choice of time origin and taken here to be zero):

$$U(t) = r_0\, e^{i\omega_0 t}\xi_0 + r_0\, e^{-i\omega_0 t}\bar{\xi}_0 + \Phi(\lambda, r_0\, e^{i\omega_0 t}\xi_0 + r_0\, e^{-i\omega_0 t}\bar{\xi}_0). \tag{35}$$

The structure of the solution obviously indicates its axisymmetric nature. Every rotation about oz,

$$R_{oz}(\theta) = e^{-i\theta J_3},$$

leaves the principal part of U unchanged ($R_{oz}(\theta)X = X$) because of (13). The complete invariance of U then directly follows from (7):

$$R_{oz}(\theta)U(t) = U(t).$$

Except for this subgroup of $O(3)$, the symmetry $R_{ox}(\pi)$ is the unique rotation acting as the identity on (35) (use (12) instead of (13)). Since we have

$$R_{ox}(\pi)R_{oz}(\theta) = R_{oz}(-\theta)R_{ox}(\pi),$$

we can conclude that $O(2)$ is the symmetry group of U. Hence this solution is nothing else than the axisymmetric solution of [7].

The stability of (35) is investigated by linearizing system (24) about (35) and by using the property stated in § 2.6. Let us introduce variables ρ_0, φ_0, y_j as follows:

$$x_0 = (r_0 + \rho_0)\, e^{i(\omega_0 t + \varphi_0)}, \qquad x_j = y_j\, e^{i\omega_0 t}, \quad j \neq 0. \tag{36}$$

The stability study then relies on an autonomous system that may be divided into three uncoupled subsystems. We write only the principal parts, since these are sufficient for determining the orbital stability. We obtain

$$\frac{d\rho_0}{dt} = 2\chi_r^{(1)} r_0^2 \rho_0, \qquad \frac{d\varphi_0}{dt} = 2\chi_i^{(1)} r_0 \rho_0, \tag{37}$$

$$\frac{dy_{-1}}{dt}=\left(\frac{b}{2}+c\sqrt{\frac{2}{3}}\right)r_0^2(y_{-1}+\bar{y}_1),\qquad \frac{d\bar{y}_1}{dt}=\left(\frac{\bar{b}}{2}+\bar{c}\sqrt{\frac{2}{3}}\right)r_0^2(y_{-1}+\bar{y}_1),$$

(38)

$$\frac{d\bar{y}_2}{dt}=y_{-2}\left[-\left(\frac{\bar{b}}{2}-\frac{\bar{c}}{\sqrt{6}}\right)r_0^2+\left(\bar{d}_8-\frac{\bar{d}_4}{2}-2\sqrt{\frac{2}{3}}\,\bar{d}_6+\frac{1}{\sqrt{6}}(\bar{d}_3+2\bar{d}_5)\right)r_0^4\right]$$
$$+\bar{y}_2\left[\left(\frac{\bar{b}}{2}-\frac{\bar{c}}{\sqrt{6}}\right)r_0^2+\left(2\bar{d}_8+\frac{\bar{d}_4}{2}-\sqrt{\frac{2}{3}}(\bar{d}_6+3\bar{d}_7)-\frac{1}{\sqrt{6}}(d_3+2d_5)\right)r_0^4\right],$$

(39)

$$\frac{dy_{-2}}{dt}=y_{-2}\left[\left(\frac{b}{2}-\frac{c}{\sqrt{6}}\right)r_0^2+\left(2d_8+\frac{d_4}{2}-\sqrt{\frac{2}{3}}(d_6+3d_7)-\frac{1}{\sqrt{6}}(d_3+2d_5)\right)r_0^4\right]$$
$$+\bar{y}_2\left[-\left(\frac{b}{2}-\frac{c}{\sqrt{6}}\right)r_0^2+\left(d_8-\frac{d_4}{2}-2\sqrt{\frac{2}{3}}\,d_6+\frac{1}{\sqrt{6}}(d_3+2d_5)\right)r_0^4\right].$$

We specify the fifth order terms in the expression (39) to remove the degeneracy of the eigenvalue zero present in the system truncated at third order.

The subsystem (37) generates the following eigenvalues:

$$\sigma_1=2\chi_r^{(1)}r_0^2+O(r_0^4),\qquad \sigma_2=0. \tag{40}$$

The eigenvalue $\sigma_2=0$ is the usual one related to the translational invariance in the choice of time origin for the Hopf bifurcating solution.

From system (38) we obtain the following eigenvalues:

$$\sigma_3=\left(b_r+2c_r\sqrt{\frac{2}{3}}\right)r_0^2+O(r_0^4),\qquad \sigma_4=0. \tag{41}$$

These eigenvalues are double since the system (38) may be decomposed into a system in (y_{-1}, y_1) and in its complex conjugate. The emergence of the double eigenvalue $\sigma_4=0$ just relies on the $O(3)$ invariance of F. As a matter of fact, there is a Γ-orbit of axisymmetric solutions generated by the action on the solution (35) of the rotation group. An infinitesimal rotation as written in Appendix 1 induces the emergence of y_{-1} and y_1 components. This explains classically the presence of the double zero eigenvalue in (38). It may seem contradictory that a three parameter family of axisymmetric solutions generates only a multiplicity of two; the "third" zero is actually taken into account by translational invariance.

The system (39) with $(y_{-2}, \bar{y}_2)$ components exhibits a linear operator of the following form:

$$\begin{pmatrix} A & B \\ \bar{B} & \bar{A} \end{pmatrix} \tag{42}$$

the eigenvalues of which are written for (39) as

$$\sigma_\pm=A_r\pm\sqrt{|B|^2-A_i^2}.$$

This leads to the following eigenvalues:

$$\sigma_5=\left(b_r-c_r\sqrt{\frac{2}{3}}\right)r_0^2+O(r_0^4),\qquad \sigma_6=\Delta r_0^4+O(r_0^6) \tag{43}$$

where

$$\Delta=3d_{8r}-\sqrt{6}(d_{6r}+d_{7r})+\frac{b_i-c_i\sqrt{2/3}}{b_r-c_{cr}\sqrt{2/3}}[3d_{8i}-\sqrt{6}(d_{6i}+d_{7i})]. \tag{44}$$

The reader should note the necessity of expanding F up to the fifth order to get a nonzero value for σ_6. Obviously, such an effort would have been hopeless for σ_4. The

number of zero eigenvalues actually coincides with the expected number given by Golubitsky and Stewart [7] using group-theoretic arguments.

Collecting (40), (41), and (43), we obtain three sufficient conditions for the axisymmetric branch to be stable orbitally:

$$
\begin{aligned}
&a_r - \frac{b_r}{2} - c_r\sqrt{\frac{3}{2}} < 0,\\
&b_r\sqrt{\frac{3}{2}} < c_r < -\left(\frac{b_r}{2}\right)\sqrt{\frac{3}{2}},\\
&\Delta < 0.
\end{aligned}
\tag{45}
$$

Note that the first condition means that this branch must bifurcate supercritically. Let us summarize the above results as a theorem.

THEOREM 1. *In the invariant subspace* $\{x_j = 0, |j| \neq 0\}$, *axisymmetric time-periodic solutions of* (I) *bifurcate. The* Γ *orbit of axisymmetric solutions is* (*orbitally*) *stable if the bifurcation is supercritical and if the coefficients of* F *in* (24) *satisfy three additional inequalities* (*see* (45)), *one of them bearing on a combination* Δ *of coefficients of terms of degree* 5.

Remark. Generically, conditions (45) are necessary for orbital stability.

3.2. The first rotating wave and the standing wave solution. If $x_j = 0$ for $j = -1, 0, 1$, then (18), (19) lead to nonzero monomials for F_m when

$$
\begin{aligned}
&2(p_2 - q_2) - 2(p_{-2} - q_{-2}) = m,\\
&p_2 - q_2 + p_{-2} - q_{-2} = 1;
\end{aligned}
$$

hence $F_m = 0$ for $m = -1, 0, 1$. Moreover, we have

$$
F_{-2}(\lambda, X) = x_{-2}\tilde{F}(\lambda, |x_{-2}|^2, |x_2|^2), \qquad F_2(\lambda, X) = x_2\tilde{F}(\lambda, |x_2|^2, |x_{-2}|^2),
\tag{46}
$$

because of $R_{ox}(\pi)$ invariance of the field F (see (12)). Truncated at fifth order, the system (5) reduces to

$$
\begin{aligned}
\frac{d}{dt}x_{-2} = x_{-2}[&i\omega + \mu + a(r_{-2}^2 + r_2^2) + d_1(r_{-2}^2 + r_2^2)^2 + (d_2 + \sqrt{6}d_5)r_{-2}^2 r_2^2\\
&- \{b + \sqrt{6}c + (d_4 + \sqrt{6}d_3)(r_{-2}^2 + r_2^2)\}r_2^2],\\
\frac{d}{dt}x_2 = x_2[&i\omega + \mu + a(r_{-2}^2 + r_2^2) + d_1(r_{-2}^2 + r_2^2)^2 + (d_2 + \sqrt{6}d_5)r_{-2}^2 r_2^2\\
&- \{b + \sqrt{6}c + (d_4 + \sqrt{6}d_3)(r_{-2}^2 + r_2^2)\}r_{-2}^2].
\end{aligned}
\tag{47}
$$

This structure is similar to that observed for a Hopf bifurcation in the presence of $O(2) \times SO(2)$ symmetry (see, for instance, the Couette-Taylor problem treated in [2]). A classical result then indicates that this leads to two kinds of periodic solutions: *rotating waves* for which either x_{-2} or x_2 equals zero, and *standing waves* for which $|x_{-2}| = |x_2|$.

3.2.1. First family of rotating waves. Let us consider the periodic solution of (47) such as $r_{-2} = 0$. The other possible choice $r_2 = 0$ is just a different element belonging to the same Γ-orbit. Thus, we obtain:

$$
x_{-2} = 0, \qquad x_2 = r_2\, e^{i(\omega_2 t + \varphi_2)},
\tag{48}
$$

with

$$\mu_r + a_r r_2^2 + d_{1r} r_2^4 + O(r_2^6) = 0, \qquad \omega_2 = \omega + \mu_i + a_i r_2^2 + d_{1i} r_2^4 + O(r_2^6). \tag{49}$$

Clearly, if $\tau(t)$ denotes the time shift operator, then the solution $X(t)$ given by (48) satisfies the identity (see (13))

$$R_{oz}(\theta)\tau(-2\theta/\omega_2)X = X.$$

Since $R_{oz}(\theta)$ and $\tau(t)$ commute with Φ (see (7)) we obtain as well:

$$R_{oz}(\theta)\tau(-2\theta/\omega_2)U = U. \tag{50}$$

This identity is the characteristic feature of a *rotating wave with two waves about oz.*

To study its stability, we proceed as in § 3.1. Suppressing the terms beyond the third-order we obtain (using obvious notation):

$$\begin{aligned} &\frac{d\rho_2}{dt} = 2a_r r_2^2 \rho_2, \qquad \frac{d\varphi_2}{dt} = 2a_i r_2 \rho_2, \\ &\frac{dy_{-2}}{dt} = -(b + c\sqrt{6}) r_2^2 y_{-2}, \qquad \frac{dy_0}{dt} = -2\sqrt{\frac{2}{3}}\, c r_2^2 y_0, \\ &\frac{dy_1}{dt} = 0, \qquad \frac{dy_{-1}}{dt} = -c\sqrt{6} r_2^2 y_{-1}. \end{aligned} \tag{51}$$

Hence we obtain seven simple eigenvalues:

$$\begin{aligned} &\sigma_1 = 2a_r r_2^2 + O(r_2^4), \qquad \sigma_2 = -(b + c\sqrt{6}) r_2^2 + O(r_2^4), \\ &\sigma_3 = \bar{\sigma}_2, \quad \sigma_4 = -2\sqrt{\frac{2}{3}}\, c r_2^2 + O(r_2^4), \quad \sigma_5 = \bar{\sigma}_4, \\ &\sigma_6 = -c\sqrt{6} r_2^2 + O(r_2^4), \qquad \sigma_7 = \bar{\sigma}_6. \end{aligned} \tag{52}$$

The eighth eigenvalue $\sigma_8 = 0$ is of triple multiplicity: it corresponds to the $O(3)$ invariance, since an infinitesimal rotation on (48) generates a component in y_1 (not considering the one in y_2), while the zero eigenvalue due to the choice of time origin still corresponds to rotations about *oz.*

Conditions for stability of this family may be summarized by

$$a_r < 0, \quad c_r > 0, \quad b_r + c_r\sqrt{6} > 0, \tag{53}$$

where the first condition means that the branch must bifurcate supercritically to be stable.

3.2.2. Standing waves. Let us now consider the time-periodic solution (46) such as $|x_{-2}| = |x_2|$. This yields:

$$x_2 = \tilde{r}_2\, e^{i(\tilde{\omega}_2 t + \psi_2)}, \qquad x_{-2} = \tilde{r}\, ^{i(\tilde{\omega}_2 t + \psi_{-2})} \tag{54}$$

with

$$\begin{aligned} &\mu_r + (2a_r - b_r - c_r\sqrt{6})\tilde{r}_2^2 + \chi_r^{(3)} \tilde{r}_2^4 + O(\tilde{r}_2^6) = 0, \\ &\tilde{\omega}_2 = \omega + \mu_i + (2a_i - b_i - c_i\sqrt{6})\tilde{r}_2^2 + \chi_i^{(3)} \tilde{r}_2^4 + O(\tilde{r}_2^6), \end{aligned} \tag{55}$$

where

$$\chi^{(3)} = 4d_1 + d_2 + \sqrt{6} d_5 - 2(d_4 + \sqrt{6} d_3).$$

The frequency $\tilde{\omega}_2$ differs from the previous one ω_0 of the axisymmetric solutions only at the order $\mu_r^2[O(\tilde{r}_2^4)]$. By a time shift and an appropriate rotation we can suppress the phases: $\psi_2 = \psi_{-2} = 0$ in (54). This solution remains completely invariant under the symmmetry $R_{ox}(\pi)$ and by the composition of a rotation of angle $\pi/2$ about oz with a translation in time of a half-period:

$$R_{ox}(\pi)U = U, \qquad R_{oz}(\pi/2)\tau(\pi/\tilde{\omega}_2)U = U. \tag{56}$$

We easily recognize the symmetry subgroup predicted in [7] for the standing wave. The two standing waves first predicted turn out to belong to the same group orbit.

To compute the stability we proceed as in § 3.1 by changing variables with adapted notation:

$$x_{\pm 2} = (\tilde{r}_2 + \rho_{\pm 2})\, e^{i(\tilde{\omega}_2 t + \varphi_{\pm 2})}, \qquad x_j = y_j\, e^{i\tilde{\omega}_2 t}, \quad j = 0, 1, -1. \tag{57}$$

We then obtain the following linear system:

$$\begin{aligned}
&\frac{d}{dt}\rho_{\pm 2} = 2(a_r - b_r - c_r\sqrt{6})\tilde{r}_2^2\rho_{\mp 2} + 2a_r\tilde{r}_2^2\rho_{\pm 2},\\
&\frac{d}{dt}\varphi_{\pm 2} = 2(a_i - b_i - c_i\sqrt{6})\tilde{r}_2\rho_{\mp 2} + 2a_i\tilde{r}_2\rho_{\pm 2},\\
&\frac{d}{dt}y_{-1} = b\tilde{r}_2^2 y_{-1} + b\tilde{r}_2^2\bar{y}_1,\\
&\frac{d}{dt}\bar{y}_1 = \bar{b}\tilde{r}_2^2 y_{-1} + \bar{b}\tilde{r}_2^2\bar{y}_1,\\
&\frac{d}{dt}y_0 = \left[\left(b - c\sqrt{\frac{2}{3}}\right)\tilde{r}_2^2 + A\tilde{r}_2^4\right]y_0 + \left[-\left(b - c\sqrt{\frac{2}{3}}\right)\tilde{r}_2^2 + B\tilde{r}_2^4\right]\bar{y}_0
\end{aligned} \tag{58}$$

with

$$A = 2\left(d_4 + 2\sqrt{6}d_6 + 2\sqrt{\frac{2}{3}}\,d_7 - \sqrt{\frac{2}{3}}\,d_3 - 2\sqrt{\frac{2}{3}}\,d_5 - 4d_8\right),$$

$$B = -A - 4(3d_8 - \sqrt{6}(d_6 + d_7)).$$

Hence the eigenvalues of (58) are the following:

$$\begin{aligned}
&\sigma_1 = 2(2a_r - b_r - c_r\sqrt{6})\tilde{r}_2^2 + O(\tilde{r}_2^4), \qquad \sigma_2 = 2(b_r + c_r\sqrt{6})\tilde{r}_2^2) + O(\tilde{r}_2^4),\\
&\sigma_3 = 2b_r\tilde{r}_2^2 + O(\tilde{r}_2^4) \quad \text{(double)}, \qquad \sigma_4 = 2\left(b_r - c_r\sqrt{\frac{2}{3}}\right)\tilde{r}_2^2 + O(\tilde{r}_2^4),\\
&\sigma_5 = -4\Delta\tilde{r}_2^4 + O(\tilde{r}_2^6) \quad (\Delta \text{ defined in § 3.1}), \qquad \sigma_6 = 0 \quad \text{(quadruple)}.
\end{aligned} \tag{59}$$

The last eigenvalue is quadruple because of dependency of orbits on four parameters: three of them originate from the action of the orthogonal group, and the remaining one from the arbitrary choice of the temporal phase. Furthermore, we can easily see that an infinitesimal rotation acting on (54) does not generate a y_0 component. This

solution is then stable if

$$
\begin{aligned}
&2a_r < b_r + c_r\sqrt{6} \quad \text{(supercritical bifurcation)},\\
&b_r < \min\left(-c_r\sqrt{6},\, c_r\sqrt{\frac{2}{3}},\, 0\right),\\
&\Delta > 0.
\end{aligned}
\tag{60}
$$

Let us summarize the results of § 3.2 as a theorem.

THEOREM 2. *In the invariant four-dimensional subspace* $\{x_1 = x_{-1} = x_0 = 0\}$ *the subsystem* (47) *satisfies an* $O(2)\times SO(2)$ *equivariance that leads to the two classical types of bifurcating time-periodic solutions: rotating waves and standing waves* (*see* [2]). *The* Γ*-orbits of these two solutions are* (*orbitally*) *stable if* (i) *they bifurcate supercritically*; (ii) *the usual conditions for stability on cubic coefficients of F in* (24) *hold in the four-dimensional subspace, and* (iii) *one additional inequality holds for rotating waves* (*see* (53)) *while three additional inequalities hold for standing waves* (*see* (60)), *one of them depending on the combination* Δ (*see* (44)) *of fifth-order terms of F.*

3.3. Second family of rotating waves. If $x_j = 0$ for $j = -2, 0, 2$ then (18), (19) lead to nonzero monomials for F_m when

$$p_1 - q_1 - p_{-1} + q_{-1} = m, \qquad p_1 - q_1 + p_{-1} - q_{-1} = 1.$$

This implies $F_m = 0$ for $m = -2, 0, 2$, and (up to "flat terms") as for (46):

$$
\begin{aligned}
F_{-1}(\lambda, X) &= x_{-1}\tilde{F}(\lambda, |x_{-1}|^2, |x_1|^2,\\
F_1(\lambda, x) &= x_1\tilde{F}(\lambda, |x_1|^2, |x_{-1}|^2).
\end{aligned}
\tag{61}
$$

Hence, truncated at third order, the system (5) reduces to

$$
\begin{aligned}
\frac{d}{dt}x_{-1} &= x_{-1}\left[i\omega + \mu + \left(a - c\sqrt{\frac{3}{2}}\right)r_{-1}^2 + \left(a - b - c\sqrt{\frac{3}{2}}\right)r_1^2\right],\\
\frac{d}{dt}x_1 &= x_1\left[i\omega + \mu + \left(a - b - c\sqrt{\frac{3}{2}}\right)r_{-1}^2 + \left(a - c\sqrt{\frac{3}{2}}\right)r_1^2\right].
\end{aligned}
\tag{62}
$$

We can proceed as in § 3.2. We obtain the following rotating wave solution:

$$x_{-1} = 0, \qquad x_1 = r_1\, e^{i(\omega_1 t + \psi_1)}, \tag{63}$$

with

$$
\begin{aligned}
&\mu_r + \left(a_r - c_r\sqrt{\frac{3}{2}}\right)r_1^2 + O(r_1^4) = 0,\\
&\omega_1 = \omega + \mu_i + \left(a_i - c_i\sqrt{\frac{3}{2}}\right)r_1^2 + O(r_1^4).
\end{aligned}
\tag{64}
$$

By the same reasoning used in 3.2.1 we can show that

$$R_{oz}(\theta)\tau(-\theta/\omega_1)U = U. \tag{65}$$

This invariance is different from that of the rotating wave of the first kind since (65) implies that this new wave rotates approximately twice as rapidly as the previous one ($\omega_1 - \omega_2 = 0(\mu_r)$).

The stability is governed by the following system (truncated at cubic terms):

$$\frac{d}{dt}\rho_1 = 2\left(a_r - c_r\sqrt{\frac{3}{2}}\right)r_1^2\rho_1, \qquad \frac{d}{dt}\varphi_1 = 2\left(a_i - c_i\sqrt{\frac{3}{2}}\right)r_1\rho_1,$$

$$\frac{d}{dt}y_{-1} = -br_1^2 y_{-1}, \qquad \frac{d}{dt}y_2 = -c\sqrt{\frac{3}{2}}\,r_1^2 y_{-2}, \tag{66}$$

$$\frac{d}{dt}y_0 = c\sqrt{\frac{2}{3}}\,r_1^2 y_0 + cr_1^2\bar{y}_2, \qquad \frac{d}{dt}\bar{y}_2 = \bar{c}\sqrt{\frac{3}{2}}\,r_1^2\bar{y}_2 + \bar{c}r_1^2 y_0,$$

from which the following eigenvalues can be exhibited:

$$\sigma_1 = 2\left(a_r - c_r\sqrt{\frac{3}{2}}\right)r_1^2 + O(r_1^4), \qquad \sigma_2 = -br_1^2 + O(r_1^4),$$

$$\sigma_3 = \bar{\sigma}_2, \quad \sigma_4 = -c\sqrt{\frac{3}{2}}\,r_1^2 + O(r_1^4), \quad \sigma_5 = \bar{\sigma}_4, \tag{67}$$

$$\sigma_6 = \left(c\sqrt{\frac{2}{3}} + \bar{c}\sqrt{\frac{3}{2}}\right)r_1^2 + O(r_1^4), \quad \sigma_7 = \bar{\sigma}_6, \quad \sigma_8 = 0 \quad \text{(triple)}.$$

For the reason already stated, the zero eigenvalue of the rotating bifurcated branch is triple. This branch, however, is generically *unstable* (if c_r is nonzero).

Now it is natural to investigate the standing wave solution given by

$$x_{-1} = \tilde{r}_1\, e^{i(\tilde{\omega}_1 t + \psi_{-1})}, \qquad x_1 = \tilde{r}_1\, e^{i(\tilde{\omega}_1 t + \psi_1)},$$

with

$$\mu_r + (2a_r - b_r - c_r\sqrt{6})\tilde{r}_1^2 + O(\tilde{r}_1^4) = 0, \qquad \tilde{\omega}_1 = \omega + \mu_i + (2a_i - b_i - c_i\sqrt{6})\tilde{r}_1^2 + O(\tilde{r}_1^4). \tag{68}$$

This solution actually belongs to the family spanned by the rotating wave (54) found in § 3.2.2 since the rotation transforms (see Appendix 1) $X = (\tilde{r}_2, 0, 0, 0, \tilde{r}_2)$ into $X' = (0, i\tilde{r}_2, 0, i\tilde{r}_2, 0)$. We can now sum up these results as a theorem.

THEOREM 3. *In the invariant four-dimensional subspace* $\{x_{-1} = x_0 = x_2 = 0\}$ *the subsystem* (62) *satisfies an* $O(2)\times SO(2)$ *equivariance as in Theorem* 2. *The rotating waves rotate approximately twice as rapidly as the ones of Theorem* 2. *Their* Γ *orbit is always unstable. Standing waves belong to the same* Γ*-orbit as those of Theorem* 2.

3.4. Tetrahedral waves. If $x_j = 0$ for $j = -1, 0, 2$ then (18), (19) lead to nonzero monomials for F_m when

$$p_1 - q_1 - 2(p_{-2} - q_{-2}) = m, \qquad p_1 - q_1 + p_{-2} - q_{-2} = 1.$$

This implies $F_m = 0$ for $m = -1, 0, 2$ and (up to "flat terms") as in (46):

$$F_{-2}(\lambda, x) = x_{-2}\tilde{F}_{-2}(\lambda, |x_{-2}|^2, |x_1|^2),$$
$$F_1(\lambda, x) = x_1\tilde{F}_1(\lambda, |x_{-2}|^2, |x_1|^2). \tag{69}$$

Furthermore, equivariance of F with respect to a rotation such as (see Appendix 1):

$$\theta = \alpha, \quad \varphi_1 = -\pi/2, \quad \varphi_2 = -\pi/2, \tag{70}$$

where

$$\sin\alpha = \frac{2}{3}\sqrt{2}, \qquad \cos\alpha = -\frac{1}{3},$$

leads to the following condition for F defined by (69):

$$\tilde{F}_1(\lambda, |x_{-2}|^2, 2|x_{-2}|^2) = \tilde{F}_{-2}(\lambda, |x_{-2}|^2, 2|x_{-2}|^2) \tag{71}$$

due to the fact that the rotation (70) leaves invariant the vector

$$X=(1,0,0,\sqrt{2},0).$$

Suppressing terms higher than sixth order, we have that the governing system takes the following form:

$$\begin{aligned}\frac{d}{dt}x_{-2}&=x_{-2}\Big[i\omega+\mu+ar_{-2}^2+(a-c\sqrt{6})r_1^2+d_1(r_1^2+r_{-2}^2)^2\\&\qquad-\sqrt{6}d_3r_1^2(r_1^2+r_{-2}^2)+3\sqrt{\frac{3}{2}}d_6r_1^4\Big],\\\frac{d}{dt}x_1&=x_1\Big[i\omega+\mu+(a-c\sqrt{6})r_{-2}^2+\Big(a-c\sqrt{\frac{3}{2}}\Big)r_1^2+d_1(r_1^2+r_2^2)^2\\&\qquad-\sqrt{\frac{3}{2}}d_3(r_1^2+r_2^2)(r_1^2+2r_{-2}^2)+3\sqrt{6}d_6r_1^2r_{-2}^2+\frac{d_9}{2}(r_1^2-2r_{-2}^2)^2\Big]\end{aligned}\tag{72}$$

where (71) is clearly satisfied.

System (72) possesses three different kinds of periodic solutions: those rotating waves found in §§ 3.2.1 and 3.3, and a solution such as

$$|x_1|=\sqrt{2}|x_{-2}|.\tag{73}$$

This relation is then valid up to arbitrary order, as can be proved by (71). Thus we obtain

$$\begin{aligned}x_1&=\sqrt{2}\hat{r}\,e^{i(\hat{\omega}t+\psi_1)},\\x_{-2}&=\hat{r}\,e^{i(\hat{\omega}t+\psi_{-2})},\end{aligned}\tag{74}$$

with

$$\begin{aligned}&\mu_r+(3a_r-2c_r\sqrt{6})\hat{r}^2+[9d_{1r}+6\sqrt{6}(d_{6r}-d_{3r})]\hat{r}^4+O(\hat{r}^6)=0,\\&\hat{\omega}=\omega+\mu_i+(3a_i-2c_i\sqrt{6})\hat{r}^2+[9d_{1i}+6\sqrt{6}(d_{6i}-d_{3i})]\hat{r}^4+O(\hat{r}^6).\end{aligned}\tag{75}$$

The phases ψ_1, ψ_{-2} may be eliminated by an appropriate rotation composed with an adequate choice of the time origin. Note that the solution $x_{-1}=\sqrt{2}x_2$, $x_j=0$ for $j=-2, 0, 1$ belongs to the same torus of solutions as (74): just rotate the angle π about ox.

The symmetry of the solution with $\psi_{-2}=\psi_1=0$ is of the form expected from [7]: a tetrahedral type solution. This means that once a rotation of angle $2\pi/3$ about oz is combined with a time translation of the third of a period, the principal part X and therefore U itself remain invariant. More explicitly, (43) implies

$$R_{oz}(2\pi/3)X=j^2X,\qquad j=e^{2i\pi/3},$$

and since

$$\tau(2\pi/3\hat{\omega})X=jX,$$

it is easy to deduce

$$R_{oz}(2\pi/3)\tau(2\pi/3\hat{\omega})U=U.\tag{76}$$

The remaining three ternary axes of this solution are shown in Fig. 1: $\mathbf{k}_1$ is located along oz; $\mathbf{k}_2$ belongs to the xoz plane and makes an angle α with $\mathbf{k}_1$ (α is defined in (70)). The axes $\mathbf{k}_3$ and $\mathbf{k}_4$ are obtained from $\mathbf{k}_2$ by rotation of $2\pi/3$ about $\mathbf{k}_1$. The solution U remains invariant under the combination of a rotation of angle $2\pi/3$ about

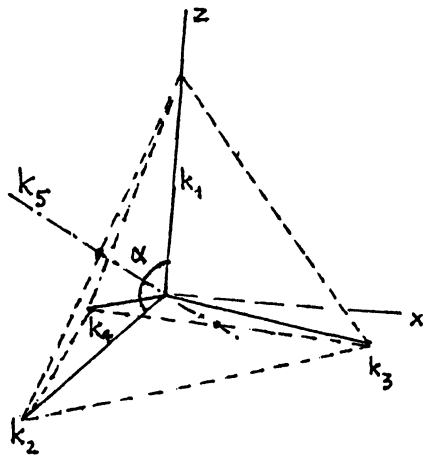


FIG. 1. *Ternary symmetry axes* (k_1, k_2, k_3, k_4) *of the tetrahedral waves.*

the axis k_i followed by a time translation of a third of the period. Let us introduce the rotation G_i that transforms $\mathbf{k}_1$ into $\mathbf{k}_i$:

$$G_i = R_{oz}(-\varphi_i)R_{ox}(-\alpha), \tag{77}$$

with

$$\varphi_2 = -\pi/2, \quad \varphi_3 = -\pi/2 - 2\pi/3, \quad \varphi_4 = -\pi/2 - 4\pi/3. \tag{78}$$

Then the rotation of angle $2\pi/3$ about $\mathbf{k}_i$ takes the following form:

$$G_i' = G_i R_{oz}(2\pi/3)G_i^{-1}. \tag{79}$$

It is clear that G_i' acts like $R_{oz}(2\pi/3)$ on $X = (x_{-2}, 0, 0, \sqrt{2}x_{-2}, 0)$. The bisectors of the ternary axes correspond to binary axes: for instance, the bisector of $(\mathbf{k}_1, \mathbf{k}_2)$ denoted by $\mathbf{k}_5$ is such that the symmetry

$$R_{k_5}(\pi) = G'' R_{oz}(\pi) G''^{-1}, \qquad \text{where } G'' = R_{oz}(\pi/2)R_{ox}(-\alpha/2) \tag{80}$$

leaves X invariant. This concludes the symmetry analysis and shows that this solution is the tetrahedral one predicted in [7].

The study of the stability yields (using obvious notation):

$$\begin{aligned}
\frac{d}{dt}\rho_{-2} &= [2a_r\rho_{-2} + 2\sqrt{2}(a_r - c_r\sqrt{6})\rho_1]\hat{r}^2, \\
\frac{d}{dt}\rho_1 &= [2\sqrt{2}(a_r - c_r\sqrt{6})\rho_{-2} + 2(2a_r - c_r\sqrt{6})\rho_1]\hat{r}^2, \\
\frac{d}{dt}\bar{y}_0 &= [\bar{c}\sqrt{6}\bar{y}_0 + \bar{c}\sqrt{2}y_{-1} + 2\bar{c}y_2]\hat{r}^2, \\
\frac{d}{dt}y_{-1} &= [c\sqrt{2}\bar{y}_0 + (-2b + c\sqrt{6})y_{-1} + b\sqrt{2}y_2]\hat{r}^2, \\
\frac{d}{dt}y_2 &= [2c\bar{y}_0 + b\sqrt{2}y_{-1} + (c\sqrt{6} - b)y_2]\hat{r}^2,
\end{aligned} \tag{81}$$

where equations for phases are omitted. The eigenvalues are now:

$$
\begin{aligned}
&\sigma_1 = 2(3a_r - 2c_r\sqrt{6})\hat{r}^2 + O(\hat{r}^4),\\
&\sigma_2 = 2c_r\sqrt{6}\hat{r}^2 + O(\hat{r}^4) \quad \text{(triple at this order)},\\
&\sigma_3 = (-3b + c\sqrt{6})\hat{r}^2 + O(\hat{r}^4),\\
&\sigma_4 = \bar{\sigma}_3,\\
&\sigma_5 = 0 \quad \text{(quadruple)}.
\end{aligned} \tag{82}
$$

The eigenvalue σ_5 is quadruple because of orbital stability (as for standing waves). The others yield the following conditions for stability:

$$
\begin{aligned}
&3a_r - 2c_r\sqrt{6} < 0 \quad \text{(supercritical bifurcation)},\\
&c_r < \min : \left(b_r \sqrt{\frac{3}{2}}; 0 \right).
\end{aligned} \tag{83}
$$

Let us sum up these results as a theorem.

THEOREM 4. *In the invariant four-dimensional subspace* $\{x_{-1} = x_0 = x_2 = 0\}$ *the subsystem* (72) *possesses three types of time-periodic bifurcating solutions*: (i) *the two different families of rotating waves found in Theorems* 2 *and* 3, *and* (ii) *a solution such that* $|x_1| = \sqrt{2}|x_{-2}|$, *which possesses the tetrahedral symmetry indicated by Golubitsky and Stewart in* [7]. *The* Γ *orbit of this solution is stable if the bifurcation is supercritical and two additional inequalities are realized on the cubic terms of F* (*see* (83)).

3.5. Summary of results. In the plane (b_r, c_r) Fig. 2 shows the domains of stability of the various periodic solutions depending on the sign of a_r. Despite the relatively large number of solutions, Fig. 2 is rather simple and the various domains just overlap on a small part of this plane (for a_r negative). One condition of stability is the usual result of supercriticality; nonetheless, this condition is not sufficient since the second family of rotating waves is proved to be always unstable. The first family of rotating waves and axisymmetric solutions may be simultaneously stable if Δ and a_r are negative.

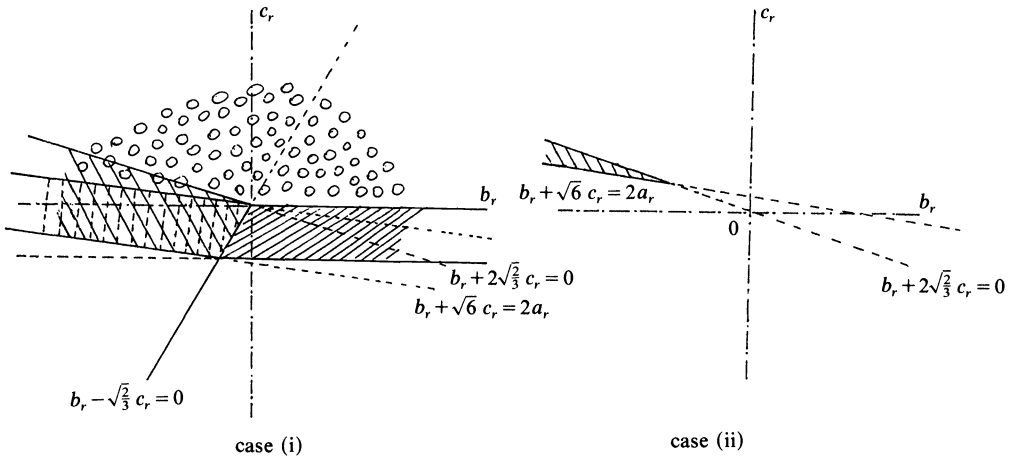


FIG. 2. *Stability domains of the periodic solutions*: *case* (i) $a_r < 0$, *case* (ii) $a_r > 0$. /// *tetrahedral waves*; \\\ *axisymmetric solution if* $\Delta < 0$; //// *standing waves of* $\Delta > 0$ (*axisymmetric solution if* $\Delta < 0$); °°° *rotating waves* (*first family*).

This is not the case for standing waves or axisymmetric waves because of Δ. Note that the fifth order terms appearing in (5) with nine coefficients play a role in the stability study by means of the single combination Δ of three of them (d_6, d_7, d_8). Finally we note that pictures of all five waves are given in Montaldi, Roberts, and Stewart [16].

4. Quasiperiodic solutions. Here we look for the possiblity of bifurcations leading to quasiperiodic patterns. As a matter of fact, the differential system may exhibit such behavior at the onset of instability since it is eight-dimensional: five amplitudes r_m and three phases $\theta_1, \theta_2, \theta_3$ appear in (5) which may be rewritten as:

$$\begin{aligned} \frac{dr_m}{dt} &= f_m(r_{-2}, r_{-1}, r_0, r_1, r_2, \theta_1, \theta_2, \theta_3), \\ \frac{d\psi_m}{dt} &= g_m(r_{-2}, r_{-1}, r_0, r_1, r_2, \theta_1, \theta_2, \theta_3), \end{aligned} \tag{84}$$

where

$$\theta_1 = 2\psi_0 - \psi_1 - \psi_{-1}, \qquad \theta_2 = 2\psi_0 - \psi_2 - \psi_{-2},$$
$$\theta_3 = 2\psi_{-1} - \psi_0 - \psi_{-2}.$$

Steady solutions $\theta_1, \theta_2, \theta_3, r_m (\neq 0)$ may be found in (84) if

$$\begin{aligned} &f_m(r_{-2}, r_{-1}, r_0, r_1, r_2, \theta_1, \theta_2, \theta_3) = 0, \\ &g_m(r_{-2}, r_{-1}, r_0, r_1, r_2, \theta_1, \theta_2, \theta_3) = \omega_m, \qquad m = -2, \cdots, 2, \end{aligned} \tag{85}$$

where

$$2\omega_0 - \omega_1 - \omega_{-1} = 2\omega_0 - \omega_2 - \omega_{-2} = 2\omega_{-1} - \omega_0 - \omega_{-2} = 0,$$

all the ω_m being close to $\omega + \mu_i$. These pulsations satisfy the relations (29) for a pair (α, β). If α and β are not rationally related, the solution is quasiperiodic with two basic frequencies β (close to ω) and α very small. Unfortunately, it is hopeless to consider the full system (84)—or (85)—since it is much too complicated for us to find analytical solutions. One method for avoiding such difficulties is to consider special cases. Systems (47), (62), and (72) have a feature in common with the $O(2) \times SO(2)$ invariant problems where a Hopf bifurcation is present since if, respectively, $b_r + c_r\sqrt{6} = 0$, $b_r = 0$, or $c_r = 0$, it is impossible to determine at cubic order which are the observable time-periodic solutions. (The standing as well as the tetrahedral waves cannot be computed). Such situations are common for codimension 2 problems. They can be treated as in the Couette-Taylor example [4], [13]. Higher-order terms (fifth and seventh order) must be computed to describe the form and stability of these solutions. We will not discuss these cases here because we only consider codimension 1 problems.

Another possibility relies on the analysis of a lower-dimensional subsystem. For instance, requiring the x_1, x_{-1} amplitudes to be zero will lead to the reduction of the system (84) to the following:

$$\begin{aligned} \frac{dr_m}{dt} &= f_m(r_{-2}, r_0, r_2, \theta), \qquad m = -2, 0, 2, \\ \frac{d\psi_m}{dt} &= g_m(r_{-2}, r_0, r_2, \theta), \end{aligned} \tag{86}$$

where

$$\theta = 2\psi_0 - \psi_{-2} - \psi_2.$$

Hence, every set $\{r_{-2}, r_0, r_2 \text{ (all} \neq 0), \theta\}$ satisfying

$$\begin{aligned} &f_m(r_{-2}, r_0, r_2, \theta) = 0, \qquad m = -2, 0, 2, \\ &2g_0 - g_{-2} - g_2 = 0, \end{aligned} \tag{87}$$

where $g_2(r, \theta)$ is different from $g_{-2}(r, \theta)$, corresponds to a quasiperiodic solution of (86) with two basic frequencies. Clearly, r_2 must differ from r_{-2}; if not, we would show via (23) that $f_{-2}(r, \theta)$ (respectively, $g_{-2}) = f_2$ (respectively g_2), which means that the solution is periodic.

System (87) truncated at the third order takes the following form:

$$\begin{aligned} &r_{-2}\left[\mu_r + a_r r_{-2}^2 + \left(a_r - 2\sqrt{\frac{2}{3}}\, c_r\right) r_0^2 + (a_r - b_r - c_r\sqrt{6}) r_2^2\right] \\ &\quad + \frac{1}{2} r_0^2 r_2 \left[\left(-b_r + c_r\sqrt{\frac{3}{2}}\right)\cos\theta - \left(-b_i + c_i\sqrt{\frac{3}{2}}\right)\sin\theta\right] = 0, \\ &r_2\left[\mu_r + \left(a_r - b_r - c_r\sqrt{6}\right) r_{-2}^2 + \left(a_r - 2\sqrt{\frac{2}{3}}\, c_r\right) r_0^2 + a_r r_2^2\right] \\ &\quad + \frac{1}{2} r_0^2 r_{-2} \left[\left(-b_r + c_r\sqrt{\frac{3}{2}}\right)\cos\theta - \left(-b_i + c_i\sqrt{\frac{3}{2}}\right)\sin\theta\right] = 0, \\ &\mu_r + \left(a_r - 2\sqrt{\frac{2}{3}}\, c_r\right)(r_{-2}^2 + r_2^2) + \frac{1}{2}(2a_r - b_r - c_r\sqrt{6}) r_0^2 \\ &\quad + r_2 r_{-2}\left[\left(-b_r + c_r\sqrt{\frac{3}{2}}\right)\cos\theta + \left(-b_i + c_i\sqrt{\frac{3}{2}}\right)\sin\theta\right] = 0, \\ &\left(-b_i + c_i\sqrt{\frac{2}{3}}\right)(r_0^2 - r_2^2 - r_{-2}^2)(1 - \cos\theta) \\ &\quad - \left(-b_i + c_i\sqrt{\frac{2}{3}}\right)\frac{(r_2 - r_{-2})^2}{2r_2 r_{-2}}(r_0^2 + 2r_2 r_{-2})\cos\theta \\ &\quad - \left(-b_r + c_r\sqrt{\frac{2}{3}}\right)[4r_2^2 r_{-2}^2 + r_0^2(r_2^2 + r_{-2}^2)]\frac{\sin\theta}{2r_2 r_{-2}} = 0. \end{aligned} \tag{88}$$

From this, we easily draw an equation for θ alone. Once this equation is solved, we obtain r_0, r_2, r_{-2} from

$$\begin{aligned} &\mu_r + \left[a_r - 2\sqrt{\frac{2}{3}}\, c_r + a_r B(\theta)\right] r_0^2 = 0, \\ &r_2 r_{-2} = A(\theta) r_0^2, \qquad r_2^2 + r_{-2}^2 = B(\theta) r_0^2, \end{aligned} \tag{89}$$

where

$$\begin{aligned} A(\theta) &= \frac{(-b_r + c_r\sqrt{3/2})\cos\theta - (-b_i + c_i\sqrt{3/2})\sin\theta}{2(b_r + c_r\sqrt{6})}, \\ B(\theta) &= \frac{(-b_r + c_r\sqrt{2/3})(b_r + c_r\sqrt{6}) + (-b_r + c_r\sqrt{3/2})^2\cos^2\theta - (-b_i + c_i\sqrt{3/2})^2\sin^2\theta}{4\sqrt{2/3}(b_r + c_r\sqrt{6})c_r}. \end{aligned} \tag{90}$$

The equation for θ then turns out to be:

$$\left(-b_i + c_i\sqrt{\frac{2}{3}}\right)\left\{2A(\theta)[1-B(\theta)] + \cos\theta[4A^{2\prime} - B]\right\}$$
$$-\left(-b_r + c_r\sqrt{\frac{2}{3}}\right)\sin\theta[4A^2 + B] = 0. \tag{91}$$

The left-hand side of (91) is an odd polynomial function of degree three in $(\sin\theta, \cos\theta)$. This implies zero, two, four, or six solutions coupled in pairs $(\theta, \theta+\pi)$. Relations (89) show that only one element in each pair of solutions of (91) is acceptable. Let us show the existence of solutions when it is assumed that

$$-b_i + c_i\sqrt{\frac{2}{3}} = 0.$$

Under such hypothesis (91) simplifies to

$$\sin\theta(4A^2(\theta) + B(\theta)) = 0. \tag{92}$$

Only zero or π are possible choices. As a matter of fact, we can assert that π is impossible because of (89). While $\theta = 0$ is a solution provided

$$\begin{aligned} &b_r \in]-c_r\sqrt{6}, -c_r\sqrt{6}/4[\quad \text{if } c_r > 0, \\ &b_r \in]-c_r\sqrt{6}/4, -c_r\sqrt{6}[\quad \text{if } c_r < 0, \end{aligned} \tag{93}$$

for b_r in this open set, it is clear that the quasiperiodic solution still exists for $-b_i + c_i\sqrt{2/3}$ near zero. Since this condition is not related to a "classical" codimension 2 problem (indicated above) we may assert that the direct bifurcation to a quasiperiodic solution is general here in a "large" open set in the parameter space.

Acknowledgments. We thank P. Chossat for his helpful remarks and we gratefully acknowledge A. Cerezo for taking time to derive the general form of the equivariant vector field $F(\lambda, X)$ (see [1]).

Appendix 1. Representation of a finite rotation. We indicate here the representation $l = 2$ of a finite rotation correcting some misprints in [6]. In the canonical basis a rotation defined by the Euler angles $\varphi_1, \theta, \varphi_2$:

$$R(\varphi_1, \theta, \varphi_2) = R_{oz}(\varphi_2)R_{ox}(\theta)R_{oz}(\varphi_1)$$

is represented by a 5×5 matrix T such that

$$T_{mn}(\varphi_1, \theta, \varphi_2) = e^{-im\varphi_2}\, e^{-im\varphi_1} u_{mn}(\theta),$$

where $u_{mn}(\theta)$ satisfy:

$$u_{mn} = u_{nm},$$

$$u_{-2,-2}(\theta) = u_{2,2}(\theta) = u_{-2,2}(\pi+\theta) = \frac{1}{4}(1+\cos\theta)^2,$$

$$u_{-2,0}(\theta) = u_{0,2}(\theta) = -\frac{1}{2}\sqrt{\frac{3}{2}}(1-\cos^2\theta),$$

$$u_{-2,-1}(\theta) = u_{1,2}(\theta) = u_{-2,1}(\pi+\theta) = u_{-1,2}(\pi+\theta) = -\frac{i}{2}\sin\theta(1+\cos\theta),$$

$$u_{-1,-1}(\theta) = u_{1,1}(\theta) = u_{-1,1}(\pi+\theta) = \frac{1}{2}(2\cos^2\theta + \cos\theta - 1),$$

$$u_{-1,0}(\theta) = u_{0,1}(\theta) = -i\sqrt{\frac{3}{2}}\cos\theta\sin\theta,$$

$$u_{0,0}(\theta) = \frac{1}{2}(3\cos^2\theta - 1).$$

An infinitesimal rotation takes the following form:

$$\begin{pmatrix} 1+2i(\varphi_1+\varphi_2) & -i\theta & 0 & 0 \\ -i\theta & 1+i(\varphi_1+\varphi_2) & -i\sqrt{3/2}\theta & 0 \\ 0 & -i\sqrt{3/2}\theta & 1 & 0 \\ 0 & 0 & -i\sqrt{3/2}\theta & -i\theta \\ 0 & 0 & 0 & 1-2i(\varphi_1+\varphi_2) \end{pmatrix}.$$

REFERENCES

[1] A. CEREZO, Université de Nice 146, Nice, France, preprint May 1987.

[2] P. CHOSSAT AND G. IOOSS, *Primary and secondary bifurcations in the Couette-Taylor problem*, Japan J. Appl. Math., 2 (1985), pp. 27-68.

[3] P. CHOSSAT, *Bifurcation and stability of convective flows in a rotating or not rotating spherical shell*, SIAM J. Appl. Math., 37 (1979), pp. 624-647.

[4] ———, *Bifurcation secondaire de solutions quasi-périodiques dans un problème de bifurcation de Hopf invariant par symétrie $O(2)$*, Comptes Rendus Acad. Sci. Paris, 302 (1986), pp. 539-541.

[5] C. ELPHICK, E. TIRAPEGUI, M. E. BRACHET, P. COULLET, AND G. IOOSS, *A simple global characterisation for normal forms of singular vector fields*, Physica D., 29 (1987), pp. 95-127.

[6] I. M. GEL'FAND, R. A. MINLOS, AND Z. YA. SHAPIRO, *Representation of the rotation and Lorentz groups and their applications*, Pergamon Press, Oxford, Elmsford, NY, 1963.

[7] M. GOLUBITSKY AND I. STEWART, *Hopf bifurcation in the presence of symmetry*, Arch. Rational Mech. Anal., 87 (1985), pp. 107-165.

[8] M. GOLUBITSKY AND D. SCHAEFFER, *Bifurcation with $O(3)$ symmetry including applications to the Bénard problem*, Com. Pure Appl. Math., 35 (1982), pp. 81-111.

[9] G. IOOSS, *Bifurcation of maps and applications*, Mathematical Studies 36, North Holland, Amsterdam, 1979.

[10] ———, *Bifurcation and Transition to Turbulence in Hydrodynamics. Bifurcation Theory and Applications*, L. Salvadori ed., Lecture Notes in Mathematics 1057, Springer-Verlag, Berlin, New York, 1984.

[11] D. HENRY, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, New York, 1981.

[12] V. IUDOVICH, *Free convection and bifurcation*, J. Appl. Math. Mech., 31 (1967), pp. 101-111.

[13] P. LAURE, *Calcul effectif de bifurcations avec rupture de symétrie en hydrodynamique*, Thèse, Université de Nice, 1987.

[14] J. E. MARSDEN AND M. MCCRACKEN, *The Hopf Bifurcation and Its Applications*, Lecture Notes in Appl. Math. Sci., 19, Springer-Verlag, Berlin, New York, 1976.

[15] W. MILLER, *Symmetry Groups and Their Applications*, Academic Press, New York, 1972.

[16] J. MONTALDI, M. ROBERTS, AND I. STEWART, *Periodic solutions near equilibria of symmetric Hamiltonian systems*, Phil. Trans. Roy. Soc. London Ser. A, to appear.

[17] A. PROSPERETTI, *Viscous effects on perturbed spherical flows*; Quart. Appl. Math. (1967), pp. 339-352.

[18] D. RUELLE, *Bifurcation in the presence of a symmetry group*, Arch. Rational. Mech. Anal., 51 (1973), pp. 136-152.

[19] A. VANDERBAUWHEDE, *Center manifolds, normal forms, and elementary bifurcations*, Dynamics Reported, 2 (1989), to appear.

[20] R. FRIEDRICH AND H. HAKEN, *Static, wavelike, and chaotic thermal convection in spherical geometries*, Phys. Rev. A, 34 (1986), pp. 2100-2120.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 533-546, May 1989

002

BIFURCATION AND ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF A DELAY-DIFFERENTIAL EQUATION WITH DIFFUSION*

MARGARET C. MEMORY†

Abstract. A scalar delay-differential equation with diffusion term in one space dimension, where the diffusivity D is a bifurcation parameter, is considered. The center manifold theory and the method of Lyapunov-Schmidt are used to describe two bifurcations from spatially constant solutions as D decreases. By modifying the equation the order of these bifurcations can be reversed. Then the existence of a compact attractor for a class of such equations is shown and the structure of part of the attractor for the modified equation is investigated. It is known that the solutions are globally L^2-bounded; bounds on the solution operator from one intermediate space to another are constructed to obtain an attractor in the $W^{2,2}$ sense.

Key words. delay-differential equation, small diffusivity, population model, bifurcation, periodic orbit, attractor

AMS(MOS) subject classifications. 34K15, 35B32

Introduction. In this paper we consider

$$\begin{aligned} &\frac{\partial}{\partial t}u(t,x)=D\frac{\partial^2}{\partial x^2}u(t,x)-\left(\frac{\pi}{2}+\mu\right)u(t-1,x)[1+u(t,x)],\ t>0, \qquad x\in(0,\pi), \\ &\frac{\partial}{\partial x}u(t,x)=0, \qquad x=0,\pi. \end{aligned} \tag{1}$$

Throughout the paper μ will be in a neighborhood of zero.

Without space dependence, the equation becomes

$$\frac{d}{dt}u(t)=-\left(\frac{\pi}{2}+\mu\right)u(t-1)[1+u(t)], \tag{2}$$

which has been proposed by Hutchinson [10] to model certain plant-eating populations; it is obtained after a change of variables from a delayed logistic equation. The value of μ depends on the delay and the growth rate of the population.

The constant zero solution of (2) is seen to be stable for $\mu>0$ by analyzing the eigenvalues of the linearized equation. Chow and Mallet-Paret [4] use integral averaging to show that a supercritical Hopf bifurcation occurs at $\mu=0$ (where the linearized equation has a pair of eigenvalues $\pm i\pi/2$) and find the approximate form of the bifurcating periodic solution for $\mu>0$. Stech also takes (2) as an example in [13], where he uses the bifurcation function to study Hopf bifurcations for functional differential equations.

With the addition of a diffusion term, we allow the population density to vary in some domain Ω, imposing no-flux boundary conditions. We want to investigate the effects of diffusion on the long-term behavior of solutions by considering two questions.

First is the local problem. Given a solution of (2) and its stability properties, what are its stability properties as a solution of (1)? In §§ 2-4, we will build on the work of Yoshida [15] and Morita [12] to deal with this question for the zero solution and the periodic solution arising in the Hopf bifurcation. Morita shows that, for fixed

* Received by the editors June 3, 1987; accepted for publication (in revised form) June 29, 1988. This material is based on research supported by a National Science Foundation Graduate Fellowship, and on the author's Ph.D. dissertation at the Division of Applied Mathematics, Brown University, Providence, Rhode Island.

† Department of Mathematics, University of Alabama, Tuscaloosa, Alabama 35487.

$\mu > 0$, this periodic solution is unstable for D less than a certain D_0. We will show that there is a $D_1 > D_0$ at which another Hopf bifurcation from zero occurs, resulting in an unstable, spatially varying, periodic solution. We also show how to destabilize the original periodic solution (as D is decreased) before this bifurcation occurs by adding nonlinear terms in $u(t-1)$ to (1).

Second is the global problem. How can we use these local analyses to say something about the global structure of solutions? Hale has shown in [7] that if a retarded functional differential equation has a compact attractor, then the corresponding diffusion equation has, for large enough D, a compact attractor with this same structure. Here we are interested in how the attractor changes as D decreases toward zero. In §§ 5 and 6, we show the existence, for arbitrary $D > 0$, of a compact attractor for a class of equations including (1). We then consider the modification of (1) noted above, for which destabilization of the original periodic solution precedes the second Hopf bifurcation, and determine the structure of part of the attractor for this equation.

1. Previous results. The phase space for (1) is $C([-1, 0], X)$ for an appropriate Banach space X of functions from $[0, \pi]$ to $\mathbb{R}$. We write $u_t \in C([-1, 0], X)$ where $u_t(\theta)(x) = u(t+\theta, x)$, $-1 \leqq \theta \leqq 0$, $x \in [0, \pi]$. With the interpretation of "$\partial^2/\partial x^2$" depending on the choice of X, (1) becomes

$$\begin{aligned} &\frac{\partial}{\partial t} u(t, x) = D \frac{\partial^2}{\partial x^2} u(t, x) + f(\mu, u_t(\cdot)(x)), \qquad x \in (0, \pi), \\ &\frac{\partial}{\partial x} u(t, x) = 0, \qquad x = 0, \pi \end{aligned} \tag{3}$$

with initial condition $u_0(\cdot)(\cdot) \in C([-1, 0], X)$ given. Yoshida [15] proves existence and uniqueness of solutions of (3) if $X = W^{2,2}(0, \pi)$. (This is a special case of his general result for domain $\Omega \subset \mathbb{R}^n$.) In addition, he makes the change of variable $U(t, x) = 1 + u(t, x)$ to get positivity of solutions in this new variable. (We will use this in § 3.) Calling $W_N^{2,2} = \{v \in W^{2,2}(0, \pi): \partial v/\partial x = 0 \text{ at } x = 0, \pi\}$, we work in $C = C([-1, 0], W_N^{2,2})$.

First we investigate the stability of the zero solution of (1). The eigenfunctions of $-\partial^2/\partial x^2$ (considered here as a densely defined operator on $L^2(0, \pi)$) are $w_n(x) = \cos nx$ with corresponding eigenvalues n^2 for $n \geqq 0$. Linearizing (1) about zero and looking for solutions $e^{\lambda t} w_n(x)$, we obtain a sequence of characteristic equations

$$\lambda + \left(\frac{\pi}{2} + \mu\right) e^{-\lambda} + Dn^2 = 0, \qquad n \geqq 0. \tag{4.n}$$

Yoshida shows (using a result in Hale [9]) that for $\mu < 0$ and any D, all roots of all the equations (4.n) have negative real parts, so the zero solution is stable.

Note that (4.0) has a pair of roots $\pm i\pi/2$ when $\mu = 0$. (This is the characteristic equation for (2).) We can use Hale's result to show that all other solutions of (4.n) for each n have negative real parts.

Therefore if we fix D and increase μ, the origin is stable until $\mu = 0$, when it loses stability in exactly two directions as a spatially constant periodic orbit grows out of it. Yoshida proves (using the Center Manifold Theory and Chow and Mallet-Paret's averaging scheme) that this orbit is stable for μ in some interval, say $\mu \in (0, \mu_1)$.

Changing our point of view slightly, we think of μ fixed and decrease D. In terms of the population model, this means the animals move more slowly in Ω. We might therefore expect spatial inhomogeneities to persist, with solutions of (1) no longer approaching those of (2). We have already seen that if $\mu < 0$ the zero solution is stable

for any D: the origin cannot be destabilized by decreasing the diffusivity. We turn our attention to small $\mu > 0$ and the possible destabilization of the periodic orbit.

Morita considers this question for a class of problems in [12]. He treats (1) (for $\Omega \subset \mathbb{R}^n$) as an example and shows that for any $\mu > 0$ there is a $D = D(\mu)$ for which the spatially constant periodic orbit is unstable. He uses the Lyapunov-Schmidt method (see, for example, [3]) to compute explicitly one of the characteristic exponents γ of the periodic orbit as a function of μ and D. For small $\mu > 0$, he is able to choose D small enough so that $\gamma(\mu, D) > 0$.

Here we summarize Morita's results for (1), putting them in a slightly different form. First it is necessary to locate the periodic orbit arising in the Hopf bifurcation. We seek a solution $y(t, \varepsilon) = \varepsilon \cos \omega t + O(\varepsilon^2)$, where $\omega = \omega(\varepsilon)$, of (2) with $\mu = \mu(\varepsilon)$. (The phase space here is $C([-1, 0], \mathbb{R})$.) On applying the Lyapunov-Schmidt method, we obtain

$$y(t, \varepsilon) = \varepsilon \cos \omega t + \tfrac{1}{5}\varepsilon^2 \cos 2\omega t + \tfrac{1}{10}\varepsilon^2 \sin \omega t + O(\varepsilon^3),$$

$$\omega(\varepsilon) = \frac{\pi}{2} - \frac{1}{20}\varepsilon^2 + O(\varepsilon^3),$$

$$\mu(\varepsilon) = \frac{3\pi - 2}{40}\varepsilon^2 + O(\varepsilon^3).$$

In practice, we change variables $s = \omega t$ (where ω is to be determined) and look for 2π-periodic solutions. The phase space is still $C([-1, 0], \mathbb{R})$, where $v_{s,\omega}(\theta) = v(s + \omega\theta)$, $-1 \leqq \theta \leqq 0$.

The stability of $p(t, x, \varepsilon) \equiv y(t, \varepsilon)$ as a solution of (1) is investigated by linearizing (1) about p and seeking solutions $e^{\gamma t} v(t) \cos nx$, where v has the same period as p. If $v(t)$ is required to have minimal period $2\pi/\omega(\varepsilon)$, applying the Lyapunov-Schmidt method leads to an equation connecting D and the characteristic exponent γ. By scaling $\gamma = \gamma_2 \varepsilon^2 + O(\varepsilon^3)$, $D = D_2 \varepsilon^2$, we obtain a quadratic equation for γ_2. If $D_2 n^2 > \pi/20$, the roots γ_2 are negative or have negative real part. If $D_2 n^2 = \pi/20$ one root is zero, and if $0 < D_2 n^2 < \pi/20$ it becomes positive. Note that if $n = 0$ the characteristic exponents obtained are those for y as a solution of (2), that is, one trivial exponent and one negative one, so it is only necessary to consider $n > 0$. The periodic solution p must therefore be unstable for $D < D^*(\mu)$, where $\mu(\varepsilon)$ is given above and $D^*(\mu(\varepsilon)) \approx \pi\varepsilon^2/20$. It will lose stability in more directions across each of the approximate bifurcation curves in Fig. 1. (Recall $\mu = \mu(\varepsilon)$.)

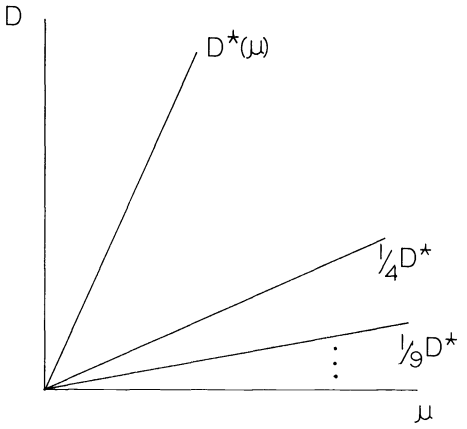


FIG. 1. *Bifurcation curves for destabilization of the periodic orbit.*

2. Second Hopf bifurcation. To study the small solutions of (1) it is certainly necessary to keep track of what is happening at the origin. For fixed $\mu > 0$ and large enough D, the origin has a two-dimensional unstable manifold. Does the origin also become less stable as D is decreased? We can see what is going on more clearly by linearizing (1) about zero and expanding in eigenfunctions $\sum_{n=0}^{\infty} u^n(t) \cos nx$. The equations decouple, and we have

$$\frac{d}{dt}u^n(t) = -Dn^2u^n(t) - \left(\frac{\pi}{2}+\mu\right)u^n(t-1) \tag{5.n}$$

for each $n \geqq 0$. Consider each (5.n) to be a separate delay-differential equation, and consider the stability of the origin as a solution of (5.n).

For the equation $\dot{u}(t) = -au(t) - bu(t-1)$, we have the situation illustrated in Fig. 2.

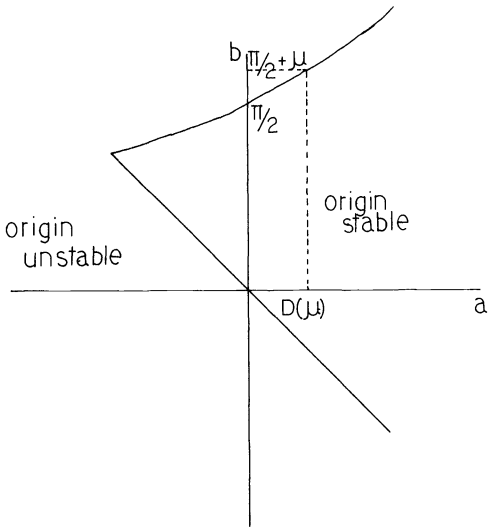


FIG. 2. *Stability of the origin.*

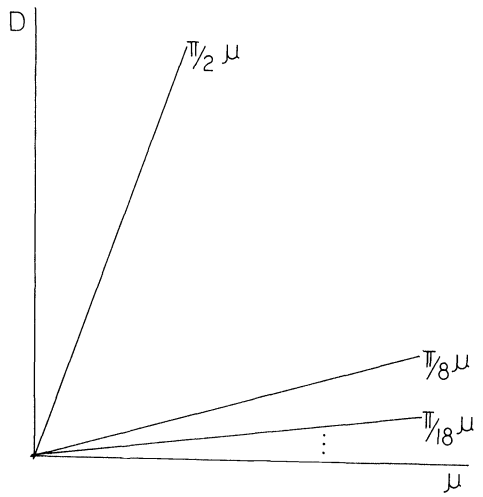


FIG. 3. *Hopf bifurcation curves.*

For (5.0) we know that the origin is unstable (with two-dimensional unstable manifold) for $\mu > 0$ and that no other bifurcations occur at the origin for the μ's we are considering.

For (5.n), $n \geqq 1$, we fix $\mu > 0$ and find $D = D(\mu)$ marked in Fig. 2, where the origin loses stability. We look for pure imaginary roots $\lambda = \pm i\beta$, $\beta > 0$, of the characteristic equation

$$\lambda + \left(\frac{\pi}{2}+\mu\right)e^{-\lambda} + Dn^2 = 0.$$

The real and imaginary parts give

$$0 = Dn^2 + \left(\frac{\pi}{2}+\mu\right)\cos\beta, \qquad \beta = \left(\frac{\pi}{2}+\mu\right)\sin\beta.$$

We expect β close to $\pi/2$, so let $\beta = \pi/2 + \tilde{\beta}$. The second equation above gives $\tilde{\beta} \approx \mu$. The first equation then gives $Dn^2 \approx \pi\tilde{\beta}/2 \approx \pi\mu/2$ for small μ.

Going back to the full equation (1), we again have a set of approximate bifurcation curves shown in Fig. 3, across each of which the already unstable origin loses stability in two more directions.

Consider the first of these bifurcations, when $D \approx \pi\mu/2$. We do not have a Hopf Bifurcation Theorem for equations for the form (1), but we can use the Center Manifold Theory to reduce the equation to a two-dimensional one where the Hopf Bifurcation Theorem holds. Yoshida [15] proves the existence of a local two-dimensional integral manifold at the origin, for fixed D and depending on μ close to μ_c where the equations (5.n) have exactly one pair of pure imaginary roots. We have been holding μ fixed and decreasing D, but the bifurcation curves are such that fixing D and increasing μ is equivalent. For any D, our $\mu_c \approx 2D/\pi$ is positive, so (5.0) has a pair of roots with positive real part. This means the center manifold will not be attracting.

Note that we could also use Carr's Center Manifold Theory for infinite-dimensional systems (see [2]), which requires a semigroup formulation of the problem slightly different from Yoshida's. Using Carr's work we get a two-dimensional center manifold depending on μ and D in a neighborhood of the bifurcation point.

Either way we know a Hopf bifurcation occurs as D decreases through $D \approx \pi\mu/2$, and we have only to determine whether a periodic orbit grows out of the origin or collapses onto it. The calculation proceeds as follows.

Substituting $u(t, x) = \sum_{n=0}^{\infty} u^n(t) \cos nx$ into (1) and calling $\alpha = \pi/2 + \mu$ gives

$$\frac{d}{dt} u^0(t) = -\alpha u^0(t-1) - \alpha u^0(t)u^0(t-1) - \frac{\alpha}{2} \sum_{n=1}^{\infty} u^n(t)u^n(t-1),$$

$$\frac{d}{dt} u^1(t) = -Du^1(t) - \alpha u^1(t-1) - \alpha u^0(t)u^1(t-1) - \alpha u^1(t)u^0(t-1)$$

$$-\frac{\alpha}{2} \sum_{n=1}^{\infty} u^{n-1}(t)u^n(t-1) - \frac{\alpha}{2} \sum_{n=2}^{\infty} u^n(t)u^{n-1}(t-1),$$

(6.p)

$$\frac{d}{dt} u^p(t) = -p^2 Du^p(t) - \alpha u^p(t-1) - \alpha u^0(t)u^p(t-1) - \alpha u^p(t)u^0(t-1)$$

$$-\frac{\alpha}{2} \sum_{n=1}^{p-1} u^{p-n}(t)u^n(t-1) - \frac{\alpha}{2} \sum_{n=1}^{\infty} u^{n+p}(t)u^n(t-1)$$

$$-\frac{\alpha}{2} \sum_{n=1}^{\infty} u^n(t)u^{n+p}(t-1), \qquad p \geqq 2.$$

We are looking for a periodic solution of the form

$$u^1(t) = b_0^1\varepsilon^2 + \varepsilon \cos \nu t + \sum_{n=2}^{\infty} b_n^1\varepsilon^2 \cos \nu nt + \sum_{n=2}^{\infty} c_n^1\varepsilon^2 \sin \nu nt + O(\varepsilon^3),$$

$$u^p(t) = b_0^p\varepsilon^2 + \sum_{n=1}^{\infty} b_n^p\varepsilon^2 \cos \nu nt + \sum_{n=1}^{\infty} c_n^p\varepsilon^2 \sin \nu nt + O(\varepsilon^3), \qquad p \neq 1,$$

with $\nu = \pi/2 + \nu_0 + \nu_1\varepsilon + \nu_2\varepsilon^2 + O(\varepsilon^3)$ for fixed small $\mu > 0$ and $D = D_0 + D_1\varepsilon + D_2\varepsilon^2 + O(\varepsilon^3)$.

As before, we rescale time $s = \nu t$ and rewrite the equations for $q^p(s) = u^p(t)$. We will substitute the desired solution into (6.p) and compare coefficients. We only need to determine the sign of D_2. (We expect the amplitude of the periodic orbit to be approximately proportional to the square root of the Hopf bifurcation parameter, so we expect $D_1 = 0$. The calculations confirm this.) To do this it is sufficient to consider

(6.0)-(6.2) only, retaining terms through $O(\varepsilon^2)$ in (6.0) and (6.2) and through $O(\varepsilon^3)$ in (6.1). In fact, we require the $O(\varepsilon^3)$ term only in the coefficients of $\sin s$ and $\cos s$ in (6.1).

Note that D_2, along with the other constants, depends on μ. Since μ is small, though, we retain only the $O(1)$ term in μ. In addition, we expect ν_0 to be small as a function of μ, so in the calculation we drop all powers of ν_0 higher than the first.

The calculation is lengthy, but we eventually find $D_2 < 0$, so a periodic orbit grows out of the origin as D is decreased (or, equivalently, as μ is increased). Note that neither the origin nor the bifurcating periodic orbit is stable, since the two-dimensional center manifold containing them is not attracting, but the origin loses its stability in the center manifold to the periodic orbit.

3. Direction of bifurcation. We now return to the bifurcation from the spatially constant periodic orbit at $D = D^*(\mu)$. Since the bifurcation occurs when a real characteristic exponent becomes positive, we expect bifurcating periodic orbits rather than an invariant torus. A calculation similar to the preceding one will allow us to determine whether the bifurcation is super-, sub-, or transcritical.

We look for the bifurcating periodic solutions

$$z(t, x, \varepsilon) = \varepsilon \cos \nu t + \varepsilon(a + b \cos \nu t + c \sin \nu t) \cos x + O(\varepsilon^2)$$

of (1) with $\mu = \mu(\varepsilon)$ and $D = D(\varepsilon)$, where $\nu = \nu(\varepsilon)$ is close to $\omega(\varepsilon)$ computed earlier. (Recall that the spatially constant periodic solution $p(t, x, \varepsilon)$ loses stability in the spatial direction $w_1(x) = \cos x$.) We substitute $z(t, x, \varepsilon)$ into (1), expanding in eigenfunctions $\cos nx$ and collecting coefficients of $\cos m\nu t$, $\sin m\nu t$. The computation is essentially an extension of the previous one, with two principal differences.

First, μ is a function of ε rather than a fixed quantity, so $\mu(\varepsilon) = \mu_2\varepsilon^2 + O(\varepsilon^3)$ must also be found. Second, the form of the periodic solution (actually, solutions) to be found is $z(t, x, \varepsilon) = \sum_{n=0}^{\infty} u^n(t) \cos nx$ where

$$u^0(t) = b_0^0\varepsilon^2 + \varepsilon \cos \nu t + \sum_{n=2}^{\infty} b_n^0\varepsilon^2 \cos \nu nt + \sum_{n=2}^{\infty} c_n^0\varepsilon^2 \sin \nu nt + O(\varepsilon^3),$$

$$u^1(t) = a\varepsilon + b_0^1\varepsilon^2 + (b\varepsilon + b_1^1\varepsilon^2 \cos \nu t) + (c\varepsilon + c_1^1\varepsilon^2) \sin \nu t$$
$$+ \sum_{n=2}^{\infty} b_n^1\varepsilon^2 \cos \nu nt + \sum_{n=2}^{\infty} c_n^1\varepsilon^2 \sin \nu nt + O(\varepsilon^3),$$

$$u^p(t) = b_0^p\varepsilon^2 + \sum_{n=1}^{\infty} b_n^p\varepsilon^2 \cos \nu nt + \sum_{n=1}^{\infty} c_n^p\varepsilon^2 \sin \nu nt + O(\varepsilon^3), \qquad p \geqq 2,$$

with $\nu(\varepsilon) = \pi/2 + \nu_2\varepsilon^2 + O(\varepsilon^3)$ and $D(\varepsilon) = D_2\varepsilon^2 + O(\varepsilon^3)$.

Here we outline the computation showing that two periodic orbits grow out of the spatially constant one as D decreases through $D^*(\mu)$. As in the original Hopf bifurcation calculation, we rescale $s = \nu t$ and call $q^n(s) = u^n(t)$.

Substituting into (1), we first collect coefficients of $\cos x$. In no case do we need to retain terms higher than $O(\varepsilon^3)$. We find that the constant term of $q^1(s) = O(\varepsilon^4)$ and, in particular, $a = b_0^1 = 0$. The $\cos 2s$ and $\sin 2s$ terms of the coefficient of $\cos x$ give b_2^1 and c_2^1 in terms of b and c. We use this information to simplify the $\cos s$ and $\sin s$ terms, which will be used later.

Next we look at coefficients of the spatially constant term, obtaining $b_0^0 = 0$ from the (time) constant term and expressions for b_2^0 and c_2^0 in terms of b and c from the

cos $2s$ and sin $2s$ terms. We use these in the cos s and sin s terms to get

$$\nu_2 = -\frac{1}{20} - \frac{3}{40} b - \frac{1}{40} c + \frac{3}{20} bc,$$

$$\mu_2 = \frac{3\pi - 2}{40}\left(1 + \frac{3}{2} b^2 + \frac{1}{2} c^2\right) + \frac{\pi + 6}{40} bc.$$

Note that if $b = c = 0$ then we get the original spatially constant periodic orbit with the proper values for μ_2 and $\nu_2 = \omega_2$.

Finally, the coefficients of cos $2x$ give $b_0^2 = 0$ from the constant term and expressions for b_2^2 and c_2^2 from the cos $2s$ and sin $2s$ terms. Returning to the cos s cos x and sin s cos x terms obtained above, we use this new informations to get

$$D_2 c = \frac{3\pi}{20} b - \frac{9\pi}{160} b^3 - \frac{7\pi}{160} b^2 c + \frac{15\pi}{160} bc^2 + \frac{\pi}{160} c^3,$$

$$D_2 b = \frac{\pi}{20} b - \frac{3\pi}{160} b^3 + \frac{21\pi}{160} b^2 c + \frac{5\pi}{160} bc^2 - \frac{3\pi}{160} c^3.$$

Now write $c = \beta b$. Multiplying the first equation by b and the second by c and subtracting, we obtain

$$b^2 = \frac{8(\beta - 3)}{3\beta^4 - 4\beta^3 - 6\beta^2 - 4\beta - 9}$$

as long as the right-hand side is defined and nonnegative. When $\beta > 3$ we will get two periodic orbits corresponding to the positive and negative choices of b.

We then find

$$D_2 = \left(\frac{1}{1 + \beta^2}\right)\left[\frac{\pi}{20} + \frac{3\pi}{20}\beta + \frac{\pi}{160} b^2(\beta^4 + 12\beta^3 - 2\beta^2 + 12\beta - 3)\right].$$

Note that $\beta = \beta_0 = 3$ gives $D_2 = \pi/20$ as expected.

Finally, we write b^2 and μ_2 as functions of β and

$$D(\beta, \mu) \approx \left(\frac{D_2(\beta)}{\mu_2(\beta)}\right)\mu$$

and compute

$$\frac{d^+}{d\beta^+}\left(\frac{D_2(\beta)}{\mu_2(\beta)}\right) < 0 \quad \text{at } \beta = 3.$$

4. Reversing the order of bifurcation. Recall that we would like to study the structure of all the small solutions of (1) as D is decreased. If ε is fixed with $\mu \approx ((3\pi - 2)/40)\varepsilon^2$, we have analyzed two bifurcations: a Hopf bifurcation from the already unstable origin at $D \approx (\pi/2)((3\pi - 2)/40)\varepsilon^2$, and a destabilization of $p(t, x, \varepsilon)$ at $D \approx \pi\varepsilon^2/20$. Note that since not all the characteristic exponents of p have been determined, we do not know if p is stable until this bifurcation point. In any case, we would like to guarantee that $p(t, x, \varepsilon)$ becomes unstable before the second Hopf bifurcation as D is decreased. We could then study a simpler structure. We will now see that by modifying (1) we can reverse the order of the two bifurcations we have studied, so that we can consider the first destabilization of p (whenever it occurs) without considering any complications caused by bifurcations from the origin.

The new equation, suppressing dependence on x, is

$$\frac{\partial}{\partial t}u(t) = D\frac{\partial^2}{\partial x^2}u(t) - \left(\frac{\pi}{2}+\mu\right)[u(t-1)+hu^3(t-1)][1+u(t)], \qquad x\in(0,\pi),$$
$$\frac{\partial u}{\partial x} = 0, \qquad x = 0, \pi, \tag{7}$$

where h is a constant to be determined. In the next section, h will be considered another parameter. Existence and uniqueness of solutions of (7) in $C([-1,0], W_N^{2,2})$ follows as in Yoshida [15], as does positivity of $U(t)=1+u(t)$.

We now extend the previous calculations to (7). The linear part of the equation is unchanged, so we still expect a Hopf bifurcation at $\mu=0$. To determine the direction of bifurcation we seek $\tilde{y}(t,\varepsilon)=\varepsilon\cos\omega t+O(\varepsilon^2)$ for new functions $\mu=\mu(\varepsilon)$ and $\omega=\omega(\varepsilon)$. We find $\omega_2=-1/20$ as before and $\tilde{y}(t,\varepsilon)=y(t,\varepsilon)+O(\varepsilon^3)$, but $\mu_2=(3\pi-2)/40-3\pi h/8$. We will consider only $h<(3\pi-2)/15\pi$ so $\mu(\varepsilon)=\mu_2\varepsilon^2+O(\varepsilon^3)>0$ for small $\varepsilon>0$ and the direction of bifurcation is unchanged.

Next we look for characteristic exponents of $\tilde{p}(t,x,\varepsilon)=\tilde{y}(t,x)$ as a solution of (7) as described in § 1. The computations are largely unchanged, and for $0\leqq h<(3\pi-2)/15\pi$ we again find one exponent becoming positive as D decreases through $\pi\varepsilon^2/20$. Note the difference here is that μ_2 is smaller for (7) than for (1): the periodic orbit destabilizes for larger D, as a function of μ, in (7) than in (1), even though D is the same as a function of ε.

The direction of bifurcation here is also unchanged. Repeating that calculation, we find that β_0, the value of β giving the single spatially constant periodic orbit at the bifurcation point, depends on h. For the values of h we want to consider (see below), $\beta<\beta_0(h)$ gives two periodic solutions, but we find in this case that $(d^-/d\beta^-)D(\beta_0(h),\mu)>0$.

Finally, we look at the second Hopf bifurcation from the origin. The bifurcation point depends only on the linear part of the equation, so for (7) also it must occur at $D(\mu)\approx\pi\mu/2$.

Repeating the computations for the direction of this bifurcation, we find that, for $h\geqq 0$, D_2 is still negative and the direction of bifurcation is the same.

We have seen that, for (7), $\tilde{p}(t,x,\varepsilon)$ undergoes bifurcation at $D=\pi\varepsilon^2/20$ and the second Hopf bifurcation from zero occurs at

$$D\approx\frac{\pi}{2}\left[\frac{3\pi-2}{40}-\frac{3\pi}{8}h\right]\varepsilon^2$$

if $h<(3\pi-2)/15\pi$. If we choose $(3\pi-6)/15\pi<h<(3\pi-2)/15\pi$, $\tilde{p}$ must lose stability first as D is decreased.

5. Existence and upper semicontinuity of the attractor. In this section we formalize the notion of looking at "all the small solutions" and show that this is possible for (7) and, as a special case, (1). We will prove the existence of a maximal compact attractor in $C=C([-1,0], W_N^{2,2})$: a compact set in C that is invariant and attracts bounded sets under the solution map and that is maximal with respect to these properties. (A set $\mathscr{A}$ attracts $\mathscr{B}$ under $S(t)$ if for any $\delta>0$ there is a t_1 such that $S(t)\mathscr{B}$ is contained in a δ-neighborhood of $\mathscr{A}$ for $t>t_1$.)

There are results on compact attractors for equations of many different types (see Hale [6] and the references therein). According to Billotti and LaSalle [1], we can prove the existence of a compact attractor for our case by showing that the solution

map is a C^0-semigroup on C, is compact for $t>1$, and is point-dissipative. ($S(t)$ is point-dissipative if there is some bounded $B\subset C$ that attracts points under $S(t)$.)

To show that the solution map is point-dissipative, we will start with a result of Luckhaus [11] for equations of the following form (suppressing dependence on x):

$$
\begin{aligned}
&\frac{\partial}{\partial t}u(t)-\frac{\partial^2}{\partial x^2}u(t)\leqq g(u(t),u(t-\tau)), \qquad x\in(0,l), \quad t>0,\\
&\frac{\partial}{\partial x}u(t)=0, \qquad x=0 \text{ and } l,\ t\geqq 0,\\
&u(t)=u_0(t), \qquad t\in[-\tau,0],\\
&u(t)\geqq 0, \qquad t\geqq 0,
\end{aligned}
\tag{8}
$$

where $g:\mathbb{R}^2\to\mathbb{R}$ satisfies

(i) $g(u,v)<u$, $u,v\in\mathbb{R}^+$;

(ii) $\lim_{\lambda\to\infty} g(\lambda u,\lambda v)/\lambda=-\infty$, $u,v\in\mathbb{R}^+$.

We will see below that (7) can be put in this form. The theorem states that, for any solution u with $u_0\in L^2[0,l]$,

$$\overline{\lim_{t\to\infty}}\,\|u(t,\cdot)\|_\infty<K$$

for some constant $K=K(g,l,\tau)$ independent of u_0. (The proof in [11] seems to give only $\overline{\lim}_{t\to\infty}\|u(t,\cdot)\|_2<K$, but we will see below that the L^∞ bound follows from the L^2 bound.) Note that this implies the solution operator is point-dissipative in $C([-\tau,0],L^\infty[0,l])$.

Based on this result, we have the following theorem.

THEOREM. *If $u_0\in L^2[0,l]$ and u is a solution of* (8) *with equality holding in the differential equation, and g satisfies* (i) *and* (ii) *and is a polynomial in u and v, then*

$$\overline{\lim_{t\to\infty}}\,\|u(t,\cdot)\|_{2,2}<K_6$$

for some constant $K_6=K_6(g,l,\tau)$ independent of u_0.

Proof. We suppose throughout that u is a solution of (8) and $\|u(t)\|_2<2K$ for $t\geqq t_0$. To simplify the notation, we take $l=\pi$.

Let $T(t)$ be the solution operator for

$$
\begin{aligned}
&\frac{\partial v}{\partial t}-\frac{\partial^2 v}{\partial x^2}=0, \qquad x\in(0,\pi), \quad t>0,\\
&\frac{\partial v}{\partial x}=0, \qquad x=0,\pi.
\end{aligned}
$$

We will get a bound on $T(t):L^2[0,\pi]\to W^{1,2}[0,\pi]$ and use the variation of constants formula to get a $W^{1,2}$ bound on u. Then a similar procedure will give the $W^{2,2}$ bound. The two steps are required to get sufficiently sharp intermediate estimates.

Let $b\in L^2(0,\pi)$ with $\partial b/\partial x=0$ at $x=0$ and π so $b=\sum_{n=2}^\infty b_n\cos nx$. Then $T(t)b=\sum_{n=0}^\infty b_n e^{-n^2t}\cos nx$, and

$$\|T(t)b\|_{1,2}^2=\sum_{n=0}^\infty b_n^2(1+n^2)\,e^{-2n^2t}\leqq\sum_{n=0}^\infty b_n^2+\sum_{n=0}^\infty n^2b_n^2\,e^{-2n^2t}.$$

For given $t>0$, we find the n for which $n^2e^{-2n^2t}$ is maximum:

$$\frac{\partial}{\partial n}(n^2e^{-2n^2t})=2ne^{-2n^2t}-4n^3te^{-2n^2t}=0$$

gives $n = 1/\sqrt{2t}$ if this is an integer. In any case $n^2 e^{-2n^2 t} \leqq e^{-1}/2t$, so we have

$$\| T(t)b \|_{1,2}^2 \leqq \left(1 + \frac{1}{2et}\right) \| b \|_2^2,$$

giving

$$\| T(t)b \|_{1,2} \leqq \frac{c}{\sqrt{t}} \| b \|_2$$

for $0 \leqq t \leqq 2\tau$, say, for some constant c.

We first use this estimate to verify the L^∞ bound on u in [11]. We can use condition (i) on g, the nonnegativity of u, and the positivity of $T(t)$ for $t \geqq 0$ to write

$$0 \leqq u(t) \leqq T(t - t_0)u(t_0) + \int_{t_0}^{t} T(t-s)u(s)\, ds,$$

pointwise in x. Since $W^{1,2}$ is continuously embedded in $C[0, \pi]$, we have for some c_1

$$\begin{aligned}
\| u(t, \cdot) \|_\infty &\leqq \| T(t - t_0)u(t_0) \|_\infty + \int_{t_0}^{t} \| T(t-s)u(s) \|_\infty \, ds \\
&\leqq c_1 \| T(t - t_0)u(t_0) \|_{1,2} + \int_{t_0}^{t} c_1 \| T(t-s)u(s) \|_{1,2} \, ds \\
&\leqq (cc_1/\sqrt{t - t_0}) \| u(t_0) \|_2 + \int_{t_0}^{t} cc_1/\sqrt{t-s} \, \| u(s) \|_2 \, ds \\
&\leqq (cc_1/\sqrt{t - t_0})(2K) + 2cc_1\sqrt{t - t_0}\,(2K) \\
&\leqq K_1 \quad \text{for } t_0 + \tau \leqq t \leqq t_0 + 2\tau.
\end{aligned}$$

The argument may be repeated with t_0 replaced by $t_0 + n\tau$ for each $n \geqq 1$, and thus, taking $K_2 = \max\{K_1, 2K\}$, we have $\| u(t) \|_2 < K_2$ and $\| u(t) \|_\infty < K_2$ for $t \geqq t_0 + \tau$.

Since $g(u(t), u(t-\tau))$ is a polynomial, we have $\| g(u(t), u(t-\tau)) \|_\infty < K_3$ and $\| g(u(t), u(t-\tau)) \|_2 < K_3$ for $t \geqq t_0 + 2\tau$ and some constant K_3. The variation of constants formula gives (using the equality in the differential equation)

$$\begin{aligned}
\| u(t) \|_{1,2} &\leqq \| T(t - (t_0 + 2\tau))u(t_0 + 2\tau) \|_{1,2} \\
&\quad + \int_{t_0 + 2\tau}^{t} \| T(t-s)g(u(s), u(s-\tau)) \|_{1,2} \, ds \\
&\leqq (c_1/\sqrt{t - (t_0 + 2\tau)})K_2 + 2c\sqrt{t - (t_0 - 2\tau)}\, K_3 \quad \text{for } t_0 + 2\tau \leqq t \leqq t_0 + 4\tau,
\end{aligned}$$

so $\| u(t) \|_{1,2} \leqq K_4$ for $t_0 + 3\tau \leqq t \leqq t_0 + 4\tau$, giving, as before, $\| u(t) \|_{1,2} \leqq K_4$ for $t \geqq t_0 + 3\tau$.

We now bound $T(t)$ as an operator from $W^{1,2}$ to $W^{2,2}$:

$$\begin{aligned}
\| T(t)b \|_{2,2}^2 &= \sum_{n=0}^{\infty} b_n^2 (1 + n^2 + n^4)\, e^{-2n^2 t} \\
&\leqq \sum_{n=0}^{\infty} b_n^2 + 2 \sum_{n=0}^{\infty} n^2 (n^2 b_n^2)\, e^{-2n^2 t}.
\end{aligned}$$

We know $\sum_{n=0}^{\infty} b_n^2 \leqq \| b \|_{1,2}^2$ and $\sum_{n=0}^{\infty} n^2 b_n^2 \leqq \| b \|_{1,2}^2$, and we have from above that $n^2 e^{-2n^2 t} \leqq 1/2et$, so

$$\| T(t)b \|_{2,2} \leqq \frac{c_2}{\sqrt{t}} \| b \|_{1,2}$$

for $0 \leqq t \leqq 2\tau$ and some c_2.

Next note that since $\|u(t)\|_{1,2} \leqq K_3$ for $t \geqq t_0+3\tau$ we have $\|g(u(t), u(t-\tau))\|_{1,2} \leqq K_5$ for $t \geqq t_0+4\tau$ and some K_5 depending on K_4 and g. As above, we use the variation of constants formula to get $\|u(t)\|_{2,2} \leqq K_6$ for $t_0+5\tau \leqq t \leqq t_0+6\tau$ and some K_6. Repeating the argument with t_0 replaced by $t_0+n\tau$ we get $\|u(t)\|_{2,2} \leqq K_6$ for $t \geqq t_0+5\tau$.

For any solution u, then, there is a t_1 such that $\|u(t)\|_{2,2} \leqq K_6$ for $t \geqq t_1$, so $\overline{\lim}_{t\to\infty} \|u(t)\|_{2,2} \leqq K_6$ independent of u_0.

We now put (7) in the form (8). Let $a = (\pi/2+\mu)(1+h)$ and rescale time $t \to t/a$. Rescale space $x \to \sqrt{D/a}\,x$. Change variables $U = U(t) = 1+u(t)$ and $V = U(t-a)$ to get

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = U\left[1 - \frac{1+3h}{1+h} V + \frac{3h}{1+h} V^2 - \frac{3h}{1+h} V^3\right], \qquad x \in (0, \sqrt{a/D}\,\pi),$$

$$\frac{\partial U}{\partial x} = 0, \qquad x = 0 \text{ and } \sqrt{a/D}\,\pi.$$

We have already noted $U \geqq 0$.

Since $-(1+3h)V + 3hV^2 - 3hV^3 < 0$ for $h \geqq 0$ and $V > 0$, hypothesis (i) on g is satisfied. It is easy to see that (ii) is satisfied also.

Next we need to show that $S(t; h, \mu)$, the solution map for (7), is a C^0-semigroup on C and is compact for $t > 1$. We proceed in two steps. First, using the work of Travis and Webb [14] we show $L(t; \mu)$, the solution map for the linearized equation

$$\frac{\partial}{\partial t} u(t) = D\frac{\partial^2}{\partial x^2} u(t) - \left(\frac{\pi}{2}+\mu\right) u(t-1), \qquad x \in (0, \pi),$$

$$\frac{\partial}{\partial x} u(t) = 0, \qquad x = 0, \pi,$$

is a C^0-semigroup on C and compact for $t > 1$. Then, using the variation of constants formula, we extend the result to $S(t; h, \mu)$.

The map $T(t)$, now considered an operator on $W_N^{2,2}$, is easily seen to have the bound $|T(t)| \leqq 1$, so by Propositions 2.1 and 3.1 in [14], $L(t; \mu)$ is a C^0-semigroup on C.

From Example 5.2 (changing the boundary condition) and Lemma 5.3 of [14], we see that $T(t)$ as an operator on $L^2(0, \pi)$ is compact for $t > 0$. Because $T(t)$ is also continuous from $L^2(0, \pi)$ to $W^{2,2}(0, \pi)$ for $t > 0$, $T(t)$ is compact on $W_N^{2,2}$ for $t > 0$. Proposition 2.4 of [14] gives $L(t; \mu)$ compact on C for $t > 1$.

From Proposition 1.1 of Yoshida [15], we know $S(t; h, \mu)$ is defined on C and continuous in t. Setting

$$g(u_t(\cdot); h, \mu) = -\left(\frac{\pi}{2}+\mu\right) u(t-1)u(t) - \left(\frac{\pi}{2}+\mu\right) hu^3(t-1)[1+u(t)],$$

we have

$$S(t; h, \mu)\phi = L(t; \mu)\phi + \int_0^t L(t-s; \mu) X_0 g(S(s; h, \mu)\phi; h, \mu)\, ds.$$

The following lemma is quoted in Yoshida [15]:

If $u, v \in W^{2,2}(0, \pi)$, then $uv \in W^{2,2}(0, \pi)$ and $\|uv\| \leqq c\|u\| \cdot \|v\|$.

Therefore $g(\cdot; h, \mu)$ maps C to $W^{2,2}(0, \pi)$ and is locally Lipschitz. An application of Gronwall's inequality, as in Corollary 2.2 of Travis and Webb [14], gives $S(t; h, \mu)\phi$

continuous in ϕ. The semigroup property clearly holds. $S(t; h, \mu)$ is compact for $t>1$ because $L(t; \mu)$ is compact.

We have shown the existence of a maximal compact attractor, which must contain any equilibrium points, periodic orbits, and orbits connecting them. The attractor for (7) thus contains all the orbits we have studied, plus any connecting orbits. It may, however, contain other invariant sets not yet identified. The structure of the full attractor is, in fact, an open problem even for (2), the equation with no space dependence.

Next we will show that the attractor for (7) does not change too abruptly as D is decreased. This will imply that, for h close to and greater than $(3\pi-6)/15\pi$ and D in a small enough range, there are no orbits connecting the spatially nonconstant periodic orbit arising in the second Hopf bifurcation from the origin and the periodic orbits arising in the destabilization of the spatially constant periodic orbit.

In practice, we will increase μ rather than decrease D. This is equivalent for the bifurcations with which we are concerned, since the bifurcation curves are nearly straight lines.

We prove that the attractor $\mathscr{A}_\mu$ for (7) (with fixed h) is upper semicontinuous in μ; that is

$$\delta(\mathscr{A}_\mu, \mathscr{A}_{\mu_0}) \to 0 \quad \text{as } \mu \to \mu_0 \text{ for each } \mu_0$$

where

$$\delta(\mathscr{A}, \mathscr{B}) = \sup_{y \in \mathscr{A}} \operatorname{dist}(y, \mathscr{B}).$$

This means that in some sense the limit of $\mathscr{A}_\mu$ is inside $\mathscr{A}_{\mu_0}$. To do this, it is enough to show (Cooperman [5], quoted in Hale [8]) that $S(t; h, \mu)$ is continuous in μ (from the variation of constants formula) and that there is some bounded $\mathscr{B} \subset C$ such that $\mathscr{A}_\mu \subset \mathscr{B}$ for all μ in a neighborhood of μ_0.

To put (7) in the form (8) this time, fix $\mu_1 > \mu_0$ and set $a = (\pi/2+\mu_1)(1+h)$ for each μ in a neighborhood of μ_0. Rescale and change variables as before to get

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = \left(\frac{\pi}{2}+\mu_1\right)^{-1}\left(\frac{\pi}{2}+\mu\right) U\left[1 - \frac{1+3h}{1+h} V + \frac{3h}{1+h} V^2 - \frac{3h}{1+h} V^3\right]$$

$$\leqq U\left[1 - \frac{1+3h}{1+h} V + \frac{3h}{1+h} V^2 - \frac{3h}{1+h} V^3\right]$$

if $\mu \leqq \mu_1$. The eventual L^2 and L^∞ bounds on U are thus independent of μ. The estimates and constructions of the other bounds above are all continuous in μ, so we can find a constant $\hat{K}$ such that

$$\overline{\lim_{t\to\infty}} |u_t(\cdot, \cdot)|_c \leqq \hat{K}$$

for any solution u of (7) with μ in a given neighborhood of μ_0. Therefore $\mathscr{A}_\mu$ is contained in the ball in C of radius $\hat{K}$ centered at the origin for each μ in this neighborhood, and $\mathscr{A}_\mu$ is upper semicontinuous at μ_0.

6. Structure of the attractor. Fix h close to and greater than $(3\pi-6)/15\pi$ and fix D. Suppose that, for (7), the bifurcation from the spatially constant periodic orbit that we have discussed occurs at $\mu = \mu^*$ and the second Hopf bifurcation occurs at $\mu = \mu^+$. Note $\mu^* < \mu^+$. The pictures below illustrate the part of the attractor we are interested in for various values of μ; they may be thought of as two-dimensional representations of a half-period map. For $\mu \leqq \mu^*$ we have the origin, the spatially

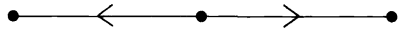


FIG. 4. *Local structure of the attractor for* $\mu \leqq \mu^*$.

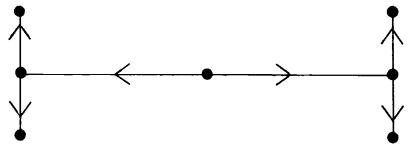


FIG. 5. *Local structure of the attractor for* $\mu^* < \mu \leqq \mu^+$.

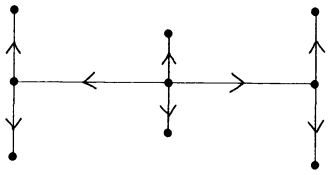


FIG. 6. *Local structure of the attractor for* $\mu > \mu^+$.

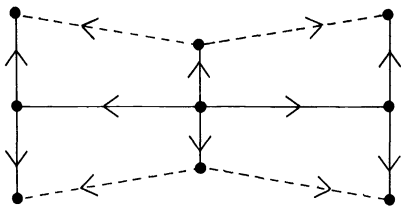


FIG. 7. *Possible connecting orbits.*

constant periodic orbit, and the orbits connecting them, which make up a two-dimensional center manifold shown in Fig. 4. For $\mu^* < \mu \leqq \mu^+$, we add the two new periodic orbits shown in Fig. 5. Note that no connections are between the origin and the new orbits, because the origin has only a two-dimensional unstable manifold, which must lie entirely in the subspace of spatially constant functions. For $\mu^+ < \mu$, we add at least the periodic orbit arising in the second Hopf bifurcation, as shown in Fig. 6.

We are interested in the existence of connecting orbits (represented in Fig. 7 by dashed lines). These connections may be formed through other bifurcations for larger μ, but they cannot exist for all small $\mu > \mu^+$, because this would violate upper semicontinuity of the attractor at μ^+.

Acknowledgments. This paper is based on my Ph.D. thesis (Division of Applied Mathematics, Brown University, 1986). The thesis was written under the direction of Jack Hale, to whom I am grateful for suggesting the problem and answering many questions. Thanks are also due to Bernold Fiedler for calling [11] to my attention, Derek Lane for discussing some estimates with me, and the referees for helpful comments.

REFERENCES

[1] J. E. BILLOTTI AND J. P. LASALLE, *Periodic dissipative processes*, Bull. Amer. Math. Soc., 77 (1971), pp. 1082-1089.

[2] J. CARR, *Applications of Centre Manifold Theory*, Springer-Verlag, Berlin, New York, 1981.

[3] S.-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, New York, 1982.

[4] S.-N. CHOW AND J. MALLET-PARET, *Integral averaging and bifurcation*, J. Differential Equations, 26 (1977), pp. 112-159.

[5] G. D. COOPERMAN, *α-condensing maps and dissipative systems*, Ph.D. thesis, Brown University, Providence, RI, June 1978.

[6] J. K. HALE, *Asymptotic behaviour and dynamics in infinite dimensions*, in Nonlinear Differential Equations, Pitman, Boston, 1985, pp. 1-42.

[7] ———, *Large diffusivity and asymptotic behavior in parabolic systems*, J. Math. Anal. Appl., 118 (1986), pp. 455-466.

[8] ———, *Some Recent Results in Dissipative Processes*, Lecture Notes in Mathematics 799, Springer-Verlag, Berlin, New York, pp. 152-172.

[9] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, New York, 1977.
[10] G. E. HUTCHINSON, *Circular causal systems in ecology*, Ann. New York Acad. Sci., 50 (1948), pp. 221-246.
[11] S. LUCKHAUS, *Global boundedness for a delay differential equation*, Trans. Amer. Math. Soc., 294 (1986), pp. 767-774.
[12] Y. MORITA, *Destabilization of periodic solutions arising in delay-diffusion systems in several space dimensions*, Japan J. Appl. Math., 1 (1984), pp. 39-65.
[13] H. W. STECH, *Hopf bifurcation calculations for functional differential equations*, J. Math. Anal. Appl., 109 (1985), pp. 472-491.
[14] C. C. TRAVIS AND G. F. WEBB, *Existence and stability for partial functional differential equations*, Trans. Amer. Math. Soc., 200 (1974), pp. 395-418.
[15] K. YOSHIDA, *The Hopf bifurcation and its stability for semilinear diffusion equations with time delay arising in ecology*, Hiroshima Math. J., 12 (1982), pp. 321-348.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 547-557, May 1989

003

A SHADOWING LEMMA WITH APPLICATIONS TO SEMILINEAR PARABOLIC EQUATIONS*

SHUI-NEE CHOW,† XIAO-BIAO LIN,‡ AND KENNETH J. PALMER§

Abstract. The property of hyperbolic sets that is embodied in the Shadowing Lemma is of great importance in the theory of dynamical systems. In this paper a new proof of the lemma is presented, which applies not only to the usual case of a diffeomorphism in finite-dimensional space but also to a sequence of possibly noninvertible maps in a Banach space. The approach is via Newton's method, the main step being the verification that a certain linear operator is invertible. At the end of the paper an application to parabolic evolution equations is given.

Key words. hyperbolic, exponential dichotomy, Shadowing Lemma, pseudo-orbit, Newton's method, parabolic evolution equation

AMS(MOS) subject classifications. 34C35, 35K22, 58F15

1. Introduction. Let f be a diffeomorphism from $\mathbf{R}^k$ into itself. Given an initial point, the iterates of f and its inverse generate a sequence of points $x_{n+1}=f(x_n)$. Then $\{x_n\}_{n\in\mathbf{Z}}$ is called the *orbit* through x_0. A sequence of points $\{y_n\}_{n\in\mathbf{Z}}$ is called a *δ-pseudo-orbit* of f if $|y_{n+1}-f(y_n)|\leqq\delta$ for all n, where $\delta>0$ is a constant. The Shadowing Lemma says that if $S\subset\mathbf{R}^k$ is a *hyperbolic* set for f then for every $\varepsilon>0$ there exists $\delta>0$ such that every δ-pseudo-orbit $\{y_n\}_{n\in\mathbf{Z}}$ in S is *ε-shadowed* by an orbit $\{x_n\}_{n\in\mathbf{Z}}$ of f, that is, $|x_n-y_n|\leqq\varepsilon$ for all n. This lemma was first stated and proved in Anosov [1] and Bowen [3] under slightly different conditions. Several different proofs were given later in Conley [4], Robinson [15], Guckenheimer, Moser, and Newhouse [6], Ekeland [5], Lanford [10], Shub [16], and Palmer [14].

A δ-pseudo-orbit can be thought of as an orbit generated numerically by a computer. If this orbit is in or near a hyperbolic set for f, the Shadowing Lemma implies that an orbit for f can be found near such a "noisy" numerical orbit for an arbitrarily long time. In fact, Hammel, Yorke, and Grebogi [7] showed how we may apply the ideas of the Shadowing Lemma to prove that "noisy" numerical orbits are actually near real orbits for a finite but fixed time even in the nonhyperbolic case. In [12], Palmer showed that the complicated behavior of the orbits of a diffeomorphism near a transversal homoclinic point can be explained by the sole use of the Shadowing Lemma. This has been generalized by Blazquez [2] to infinite-dimensional systems generated by parabolic evolution equations.

When considered abstractly, the problem of finding a shadowing orbit can be approached by Newton's method for finding zeros of functions. To see this, let X be the Banach space of all bounded $\mathbf{R}^k$-valued sequences $\mathbf{x}=\{x_n\}_{n\in\mathbf{Z}}$ with the usual sup norm and define $\mathcal{F}:X\to X$ by $(\mathcal{F}(\mathbf{x}))_n=x_n-f(x_{n-1})$, where $(\mathcal{F}(\mathbf{x}))_n$ denotes the nth element of the sequence $\mathcal{F}(\mathbf{x})\in X$. Thus $\mathbf{x}=\{x_n\}$ is an orbit of f if and only if $\mathcal{F}(\mathbf{x})=0$ and $\mathbf{y}=\{y_n\}$ is a δ-pseudo-orbit if and only if $\|\mathcal{F}(\mathbf{y})\|\leqq\delta$. The Shadowing Lemma says that if f is hyperbolic and there exists a good approximate (δ sufficiently small) solution $\mathbf{y}$ of the equation $\mathcal{F}=0$, then there exists a solution $\mathbf{x}$ near $\mathbf{y}$.

* Received by the editors February 3, 1988; accepted for publication June 22, 1988.

† Center for Dynamical Systems and Nonlinear Studies, School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332. The work of the author was partially supported by the Defense Advanced Research Projects Agency.

‡ Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695.

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

For an analogue in the continuous time case, consider the abstract ordinary differential equation in a Banach space X:

$$\dot{x} + Ax = f(x, t), \tag{1.1}$$

where A is linear and f nonlinear. A typical example of (1.1) is the nonlinear heat equation. Suppose the real line $\mathbf{R}$ is partitioned as $\mathbf{R} = \cup_{n\in\mathbf{Z}}[\tau_{n-1}, \tau_n]$ and each $x_n(t):[\tau_{n-1}, \tau_n] \to X$ is an approximate solution, that is,

$$\dot{x}_n(t) + Ax_n(t) = f(x_n(t), t) + h_n(t),$$

$$x_n(\tau_n) - x_{n+1}(\tau_n) = g_n,$$

where $h_n(t)$ and g_n are the error terms. The problem is to find an analogue of the hyperbolicity condition to guarantee that there exists a solution of (1.1) which, for each n in $\mathbf{Z}$, is close to $x_n(t)$ in the interval $[\tau_{n-1}, \tau_n]$.

In this paper, we will show that a Shadowing Lemma may be derived using Newton's method and that the lemma is applicable to the just-mentioned situation (see § 6). Because of the applications, we will work with C^1 maps in Banach spaces that are not necessarily diffeomorphisms. In fact, we consider a sequence $\{f_n\}_{n\in\mathbf{Z}}$ of mappings rather than a single mapping f. We set up the problem abstractly in a Banach space of sequences and apply a variant of Newton's method. The key tool is Lemma 3.2 in which we show that a certain linear operator is invertible (in the finite-dimensional case, this can also be proved by the perturbation theorem for exponential dichotomies in Palmer [13]). Lin has proved a similar lemma in [11], where the application is to a problem in ordinary differential equations. Here Lemma 3.2 is proved by an iteration method, which means, we believe, that it could be implemented on the computer. Newhouse [6] also used an iteration process but it is rather more involved in that at each step it uses the intersection of the stable and unstable manifolds.

Finally we should mention that as this paper was being written Walther sent us the preprint [18] where he proves the Shadowing Lemma for noninvertible maps. Stoffer [17] has also proved such a theorem. Both of these authors use the methods of Kirchgraber [9], which are quite different from ours.

2. Definition and statement of the Shadowing Lemma. What we are going to prove is a "nonautonomous" Shadowing Lemma for a sequence $f_n: X_n \to X_{n+1} (n \in \mathbf{Z})$ of C^1 maps. Here X_n is a Banach space with norm $|\cdot|_{X_n}$ (or simply $|\cdot|$ if no confusion should arise). Assume $S_n \subset X_n$, $n \in \mathbf{Z}$, is *invariant* under f_n in the sense that $f_n(S_n) \subset S_{n+1}$. Also we assume that $f_n(x)$, $Df_n(x)$ are bounded and continuous in a closed Δ-neighborhood O_n of S_n uniformly in $x \in O_n$ and $n \in \mathbf{Z}$.

We want to define what is meant by saying $\{S_n\}_{n\in\mathbf{Z}}$ is *hyperbolic*. First there is a *splitting* into closed subspaces

$$X_n = E_n^s(x) \oplus E_n^u(x) \tag{2.1}$$

for x in S_n. We require this splitting to be *invariant* in the sense that

$$Df_n(x)E_n^s(x) \subset E_{n+1}^s(f_n(x)),\ Df_n(x)E_n^u(x) \subset E_{n+1}^u(f_n(x))$$

for all x in S_n, and also *continuous*; that is, if $\mathbf{P}_n(x)$ is the projection with range $E_n^s(x)$ and nullspace $E_n^u(x)$, $\mathbf{P}_n(x)$ is continuous in the operator norm, uniformly with respect to $x \in S_n$ and $n \in \mathbf{Z}$. In terms of $\mathbf{P}_n(x)$, the invariance of the splitting is equivalent to

$$Df_n(x)\mathbf{P}_n(x) = \mathbf{P}_{n+1}(f_n(x))Df_n(x) \tag{2.2}$$

for all x in S_n. We also assume that $Df_n(x): E_n^u(x) \to E_{n+1}^u(f_n(x))$ is an isomorphism with a (bounded) inverse $(Df_n(x))^{-1}: E_{n+1}^u(f_n(x)) \to E_n^u(x)$.

Second, we require that there exist constants $K \geqq 1$, $0 \leqq \lambda < 1$ such that for any finite sequence x_m, $x_{m+1} = f_m(x_m)$, $x_{m+2} = f_{m+1}(x_{m+1}), \cdots, x_n = f_{n-1}(x_{n-1})$ with $x_m \in S_m$ and any integers $n \geqq m$,

$$
\begin{aligned}
&|Df_n(x_n)Df_{n-1}(x_{n-1}) \cdots Df_m(x_m)\mathbf{P}_m(x_m)| \leqq K\lambda^{n-m+1}, \\
&|Df_m(x_m)^{-1}Df_{m+1}(x_{m+1})^{-1} \cdots Df_n(x_n)^{-1}(I - \mathbf{P}_{n+1}(x_{n+1}))| \leqq K\lambda^{n-m+1}.
\end{aligned} \tag{2.3}
$$

Also we assume that $|\mathbf{P}_n(x)| \leqq K$, $|I - \mathbf{P}_n(x)| \leqq K$ for $x \in S_n$, $n \in \mathbf{Z}$.

An *orbit* for $\{f_n\}_{n\in\mathbf{Z}}$ is a sequence $\{x_n\}_{n\in\mathbf{Z}}$ with $x_n \in X_n$ and $x_{n+1} = f_n(x_n)$ for all $n \in \mathbf{Z}$. If $\delta > 0$ is a constant, a sequence $\{y_n\}_{n\in\mathbf{Z}}$ with $y_n \in X_n$ is said to be a *δ-pseudo-orbit* for $\{f_n\}$ if

$$|f_n(y_n) - y_{n+1}| \leqq \delta$$

for all integers n. A sequence $\{x_n\}_{n\in\mathbf{Z}}$ with $x_n \in X_n$ is said to *ε-shadow* $\{y_n\}_{n\in\mathbf{Z}}$, $y_n \in X_n$, if

$$|x_n - y_n| \leqq \varepsilon$$

for all $n \in \mathbf{Z}$.

THE SHADOWING LEMMA. *Let $\{X_n\}$, $\{f_n\}$, $\{S_n\}$, $n \in \mathbf{Z}$, be defined as above and satisfy all the properties listed above, that is,*

(i) *S_n is invariant under f_n;*

(ii) *There is a closed Δ-neighborhood O_n of S_n such that $f_n(x)$ and $Df_n(x)$ are bounded and continuous uniformly with respect to x in O_n and n in $\mathbf{Z}$;*

(iii) *$\{S_n\}_{n\in\mathbf{Z}}$ is hyperbolic.*

Then there exists $\varepsilon_0 > 0$ with the property that if $0 < \varepsilon \leqq \varepsilon_0$ there is $\delta = \delta(\varepsilon) > 0$ such that if $\{y_n\}$, $y_n \in S_n$, is a δ-pseudo-orbit for $\{f_n\}$ then there is a unique orbit $\{x_n\}$ which ε-shadows $\{y_n\}$.

To prove the Shadowing Lemma, we will use some facts about linear difference equations and a variant of Newton's method for solving nonlinear equations.

3. Facts about linear difference equations. For each integer n let $A_n : X_n \to X_{n+1}$ be a bounded linear mapping. Denote by $\Phi(n, m)(n \geqq m)$ the *transition matrix* for the *linear difference equation*

$$x_n = A_{n-1}x_{n-1}, \qquad x_n \in X_n, \quad n \in \mathbf{Z}, \tag{3.1}$$

that is,

$$\Phi(n, m) = \begin{cases} A_{n-1}A_{n-2} \cdots A_m, & n > m, \\ I, & n = m. \end{cases}$$

Equation (3.1) is said to have an *exponential dichotomy* if there is a projection valued function $P_n : X_n \to X_n$ and constants $K \geqq 1$, $0 \leqq \lambda < 1$ such that

$$\Phi(n, m)P_m = P_n\Phi(n, m) \quad \text{for } n \geqq m, \tag{3.2}$$

$$|\Phi(n, m)P_m| \leqq K\lambda^{n-m} \quad \text{for } n \geqq m. \tag{3.3}$$

Moreover, it is required that $\Phi(n, m) : \mathcal{N}(P_m) \to \mathcal{N}(P_n)$ ($\mathcal{N}$ denotes nullspace) be an isomorphism. Then for $n \geqq m$ we define $\Phi(m, n) : \mathcal{N}(P_n) \to \mathcal{N}(P_m)$ as the inverse of $\Phi(n, m) : \mathcal{N}(P_m) \to \mathcal{N}(P_n)$ and require that

$$|\Phi(m, n)(I - P_m)| \leqq K\lambda^{n-m} \quad \text{for } n \geqq m. \tag{3.4}$$

It is clear from the definition of hyperbolicity that the following lemma holds.

LEMMA 3.1. *Let $\{S_n\}_{n\in\mathbf{Z}}$ be a hyperbolic set for a sequence $f_n : X_n \to X_{n+1}$ of C^1 mappings. Then if $\{x_n\}$ is an orbit of $\{f_n\}$ with $x_n \in S_n$ for all n, the linear difference equation*

$$u_n = Df(x_{n-1})u_{n-1}$$

has an exponential dichotomy with projections $P_n = \mathbf{P}(x_n)$ *and constants* K, λ, *the projections* $\mathbf{P}(x)$ *and the constants* K, λ *being those defining the hyperbolicity of* $\{S_n\}$.

We denote by ΠX_n the Banach space of bounded sequences $\mathbf{x} = \{x_n\}_{n\in\mathbf{Z}}$, $x_n \in X_n$, with norm

$$\|\mathbf{x}\| = \|\{x_n\}\| = \sup_{n\in\mathbf{Z}} |x_n|_{X_n}.$$

If $\sup_{n\in\mathbf{Z}} |A_n| < \infty$, we can associate with the linear difference equation (3.1) the linear operator $L: \Pi X_n \to \Pi X_n$ defined by

$$(L\mathbf{x})_n = x_n - A_{n-1}x_{n-1}.$$

It turns out that if (3.1) has an exponential dichotomy then L is invertible. Now in the proof of the Shadowing Lemma we are confronted with a linear difference equation for which the existence of an exponential dichotomy is not obvious. For this reason we need the following lemma.

LEMMA 3.2. *Assume* $\sup_{n\in\mathbf{Z}} |A_n| < \infty$. *For each* $n \in \mathbf{Z}$ *let* Q_n *be a projection such that* $|Q_n| \leqq K$, $|I - Q_n| \leqq K$ *and* $|Q_{n+1}A_n(I - Q_n)| \leqq \delta$, $|(I - Q_{n+1})A_nQ_n| \leqq \delta$. *Suppose also that for all* $n \in \mathbf{Z}$ $|A_nQ_n| \leqq \lambda$ *and that* $(I - Q_{n+1})A_n : \mathcal{N}(Q_n) \to \mathcal{N}(Q_{n+1})$ *has an inverse* B_n *with* $|B_n(I - Q_{n+1})| \leqq \lambda$. *Then if* $8K\lambda \leqq 1$, $8\delta \leqq 1$ *the operator* $L: \Pi X_n \to \Pi X_n$ *defined by* $(L\mathbf{x})_n = x_n - A_{n-1}x_{n-1}$ *is invertible with* $\|L^{-1}\| \leqq 2K + 1$.

Proof. First we show L is onto. To do this we define the linear mapping $S: \Pi X_n \to \Pi X_n$ by $(S\mathbf{h})_n = Q_nh_n - B_n(I - Q_{n+1})h_{n+1}$. Then S is bounded with $\|S\| \leqq K + \lambda$ and for all n

$$\begin{aligned}
|(LS\mathbf{h})_n - h_n| &= |Q_nh_n - B_n(I - Q_{n+1})h_{n+1} - A_{n-1}\{Q_{n-1}h_{n-1} - B_{n-1}(I - Q_n)h_n\} - h_n| \\
&= |-B_n(I - Q_{n+1})h_{n+1} - A_{n-1}Q_{n-1}h_{n-1} + Q_nA_{n-1}B_{n-1}(I - Q_n)h_n| \\
&\qquad\qquad \text{since } (I - Q_n)A_{n-1}B_{n-1}(I - Q_n) = I - Q_n \\
&= |-B_n(I - Q_{n+1})h_{n+1} - A_{n-1}Q_{n-1}h_{n-1} + Q_nA_{n-1}(I - Q_{n-1})B_{n-1}(I - Q_n)h_n| \\
&\leqq |B_n(I - Q_{n+1})||h_{n+1}| + |A_{n-1}Q_{n-1}||h_{n-1}| \\
&\quad + |Q_nA_{n-1}(I - Q_{n-1})||B_{n-1}(I - Q_n)||h_n| \\
&\leqq \lambda(2 + \delta)\|\mathbf{h}\| \\
&\leqq \tfrac{1}{2}\|\mathbf{h}\|.
\end{aligned}$$

Hence $\|LS - I\| \leqq \frac{1}{2}$ and so LS has an inverse T with $\|T\| \leqq (1 - \|LS - I\|)^{-1} \leqq 2$. Then $ST = L_R^{-1}$ is a right inverse of L with

$$\|L_R^{-1}\| \leqq \|S\|\,\|T\| \leqq 2(K + \lambda) \leqq 2K + 1.$$

All that remains is to show that L is one-to-one. First note that for all $x \in X_n$,

$$\begin{aligned}
|(I - Q_n)x| &= |B_n(I - Q_{n+1})A_n(I - Q_n)x| \\
&\leqq \lambda|(I - Q_{n+1})A_n(I - Q_n)x|
\end{aligned}$$

so that

$$|(I - Q_{n+1})A_n(I - Q_n)x| \geqq \lambda^{-1}|(I - Q_n)x|. \tag{3.5}$$

Suppose L is not one-to-one. Then there exists a nonzero bounded sequence $\{x_n\}$ in ΠX_n such that $x_n = A_{n-1}x_{n-1}$ for all n. Suppose that for some n

$$|(I-Q_n)x_n| > |Q_n x_n|. \tag{3.6}$$

Then, using (3.5),

$$\begin{aligned}
|(I-Q_{n+1})x_{n+1}| - |Q_{n+1}x_{n+1}| &= |(I-Q_{n+1})A_n x_n| - |Q_{n+1}A_n x_n| \\
&\geqq |(I-Q_{n+1})A_n(I-Q_n)x_n| - |(I-Q_{n+1})A_n Q_n x_n| \\
&\quad - |Q_{n+1}A_n Q_n x_n| - |Q_{n+1}A_n(I-Q_n)x_n| \\
&\geqq \lambda^{-1}|(I-Q_n)x_n| - \delta|x_n| - K\lambda|x_n| - \delta|x_n| \\
&\geqq (\lambda^{-1} - 4\delta - 2K\lambda)|(I-Q_n)x_n| \qquad \text{since } |x_n| \leqq 2|(I-Q_n)x_n| \\
&\geqq 7|(I-Q_n)x_n|.
\end{aligned}$$

This implies that $|(I-Q_{n+1})x_{n+1}| > |Q_{n+1}x_{n+1}|$ and that $|(I-Q_{n+1})x_{n+1}| \geqq 7|(I-Q_n)x_n|$. So if (3.6) holds for some $n = m$, it holds for all $n \geqq m$ and

$$\begin{aligned}
|x_n| \geqq K^{-1}|(I-Q_n)x_n| &\geqq K^{-1}7^{n-m}|(I-Q_m)x_m| \\
&\geqq \tfrac{1}{2}K^{-1}7^{n-m}|x_m|.
\end{aligned}$$

Thus $|x_n| \to \infty$ as $n \to \infty$, contradicting the boundedness of $\{x_n\}$.

Hence it must be that $|Q_n x_n| \geqq |(I-Q_n)x_n|$ *for all* n and then

$$\begin{aligned}
|Q_{n+1}x_{n+1}| &= |Q_{n+1}A_n x_n| \\
&\leqq |Q_{n+1}A_n Q_n x_n| + |Q_{n+1}A_n(I-Q_n)x_n| \\
&\leqq (K\lambda + \delta)|x_n| \\
&\leqq 2(K\lambda + \delta)|Q_n x_n| \\
&\leqq \tfrac{1}{2}|Q_n x_n|.
\end{aligned}$$

Now there exists some m such that $Q_m x_m \neq 0$. Then for all $n \leqq m$, $|Q_n x_n| \geqq 2^{-(n-m)}|Q_m x_m| \to \infty$ as $n \to -\infty$. Again this contradicts the boundedness of $\{x_n\}$. So L must be one-to-one.

4. Newton's method for solving nonlinear equations. In this section we prove the following variant of Newton's method for solving nonlinear equations.

PROPOSITION 4.1. *Let X be a Banach space, $U \subset X$ an open subset and $\mathcal{F}: U \to X$ a C^1 mapping. Let y be a point in U such that $D\mathcal{F}(y)^{-1}$ exists and let $\varepsilon_0 > 0$ be chosen so that*

$$\|D\mathcal{F}(x) - D\mathcal{F}(y)\| \leqq (2\|D\mathcal{F}(y)^{-1}\|)^{-1} \tag{4.1}$$

for $\|x-y\| \leqq \varepsilon_0$. Then if $0 < \varepsilon \leqq \varepsilon_0$ and

$$\|\mathcal{F}(y)\| \leqq \varepsilon(2\|D\mathcal{F}(y)^{-1}\|)^{-1}, \tag{4.2}$$

the equation

$$\mathcal{F}(x) = 0 \tag{4.3}$$

has a unique solution x such that $\|x-y\| \leqq \varepsilon$.

Proof. We write

$$\mathcal{F}(x) = \mathcal{F}(y) + D\mathcal{F}(y)(x-y) + \eta(x).$$

When $\|x_1 - y\|, \|x_2 - y\| \leqq \varepsilon_0$,

$$\begin{aligned} \|\eta(x_1) - \eta(x_2)\| &= \|\mathcal{F}(x_1) - \mathcal{F}(x_2) - D\mathcal{F}(y)(x_1 - x_2)\| \\ &\leqq \left\| \int_0^1 D\mathcal{F}(x_2 + \theta(x_1 - x_2)) - D\mathcal{F}(y)\, d\theta \right\| \cdot \|x_1 - x_2\| \\ &\leqq (2\|D\mathcal{F}(y)^{-1}\|)^{-1} \|x_1 - x_2\|, \end{aligned} \tag{4.4}$$

using (4.1).

We can rewrite (4.3) as

$$x = y - D\mathcal{F}(y)^{-1}\{\mathcal{F}(y) + \eta(x)\} := T(x).$$

For $0 < \varepsilon \leqq \varepsilon_0$, we define $B_\varepsilon = \{x \in X : \|x - y\| \leqq \varepsilon\}$ and show that T is a contraction on B_ε. The proposition will then follow immediately from the contraction mapping principle.

Note first if $x \in B_\varepsilon$ then

$$\begin{aligned} \|T(x) - y\| &= \|D\mathcal{F}(y)^{-1}\{\mathcal{F}(y) + \eta(x)\}\| \\ &\leqq \|D\mathcal{F}(y)^{-1}\|\{\varepsilon(2\|D\mathcal{F}(y)^{-1}\|)^{-1} + (2\|D\mathcal{F}(y)^{-1}\|)^{-1}\|x - y\|\} \\ &= \varepsilon/2 + \|x - y\|/2 \\ &\leqq \varepsilon/2 + \varepsilon/2 = \varepsilon, \end{aligned}$$

where we have used (4.2) and (4.4) with $x_1 = x$, $x_2 = y$. Hence T maps B_ε into itself. Moreover if $x_1, x_2 \in B_\varepsilon$ then, using (4.4),

$$\begin{aligned} \|T(x_1) - T(x_2)\| &= \|D\mathcal{F}(y)^{-1}\{\eta(x_1) - \eta(x_2)\}\| \\ &\leqq \|D\mathcal{F}(y)^{-1}\| \cdot (2\|D\mathcal{F}(y)^{-1}\|)^{-1}\|x_1 - x_2\| \\ &= 1/2\|x_1 - x_2\|. \end{aligned}$$

Thus T is indeed a contraction on B_ε and the proof is completed.

5. Proof of the Shadowing Lemma. We need three lemmas for the proof.

LEMMA 5.1. *Let X be a Banach space and let $P, Q : X \to X$ be projections such that $|P|, |Q| \leqq K$. Then if $|P - Q| < 1/2K$, the operator $J = PQ + (I - P)(I - Q)$ is invertible with $|J^{-1}| \leqq (1 - 2K|P - Q|)^{-1}$. Moreover, $J(\mathcal{R}(Q)) = \mathcal{R}(P)$, $J(\mathcal{N}(Q)) = \mathcal{N}(P)$.*

Proof.

$$\begin{aligned} |J - I| &= |J - P^2 - (I - P)^2| \\ &= |P(Q - P) + (I - P)(P - Q)| \leqq 2K|P - Q| < 1. \end{aligned}$$

So J is invertible with $|J^{-1}| \leqq (1 - |J - I|)^{-1} \leqq (1 - 2K|P - Q|)^{-1}$. Clearly $J(\mathcal{R}(Q)) \subset \mathcal{R}(P)$, $J(\mathcal{N}(Q)) \subset \mathcal{N}(P)$ and equality follows from the invertibility of J. So the proof of the lemma is complete.

Now by assumption $|Df_n(x)|$ is bounded in a closed Δ-neighborhood O_n of S_n, uniformly in $x \in O_n$ and $n \in \mathbf{Z}$. Let this bound be M. Then

$$|f_n(x) - f_n(y)| \leqq M|x - y|$$

for $x \in S_n$, $y \in X_n$, and $|x - y| \leqq \Delta$. This fact is used in the following two lemmas, which make precise a statement of Guckenheimer, Moser, and Newhouse [6] that in the Shadowing Lemma it is enough to shadow a δ-pseudo-orbit for the sequence of mappings $\{f_{nk+k-1} \circ \cdots \circ f_{nk+1} \circ f_{nk}\}_{n \in \mathbf{Z}}$.

LEMMA 5.2. *If $\{y_n\}_{n\in\mathbf{Z}}$ is a δ-pseudo-orbit for $\{f_n\}$ with $y_n \in S_n$ for all n, then $\{y_{nk}\}_{n\in\mathbf{Z}}$ is a $\delta(1+M+\cdots+M^{k-1})$-pseudo-orbit for $\{f_{nk+k-1}\circ\cdots\circ f_{nk+1}\circ f_{nk}\}_{n\in\mathbf{Z}}$.*

Proof. We prove by induction that

$$|y_{nk+i}-(f_{nk+i-1}\circ\cdots\circ f_{nk+1}\circ f_{nk})(y_{nk})|\leqq\delta(1+M+\cdots+M^{i-1})$$

for $1\leqq i\leqq k$. Since $\{y_n\}$ is a δ-pseudo-orbit, it certainly holds for $i=1$. Assuming it for $i\geqq 1$, we prove it for $i+1$ as follows:

$$\begin{aligned}
&|y_{nk+i+1}-(f_{nk+i}\circ\cdots\circ f_{nk+1}\circ f_{nk})(y_{nk})|\\
&\quad\leqq|y_{nk+i+1}-f_{nk+i}(y_{nk+i})|+|f_{nk+i}(y_{nk+i})-f_{nk+i}((f_{nk+i-1}\circ\cdots\circ f_{nk})(y_{nk}))|\\
&\quad\leqq\delta+M|y_{nk+i}-(f_{nk+i-1}\circ\cdots\circ f_{nk})(y_{nk})|\\
&\quad\leqq\delta+M\delta(1+M+\cdots+M^{i-1})=\delta(1+M+\cdots+M^{i}).
\end{aligned}$$

LEMMA 5.3. *Let $\{y_n\}$ be a δ-pseudo-orbit for $\{f_n\}$ with $y_n\in S_n$, and let $\{x_n\}$ be an orbit of $\{f_n\}$ such that $\{x_{nk}\}\varepsilon$-shadows $\{y_{nk}\}$, $k\geqq 1$ being fixed. Set $\varepsilon_1=\max\{\varepsilon,\delta\}$. Then if $\varepsilon_1(1+M+\cdots+M^k)\leqq\Delta$, $\{x_n\}$ $\varepsilon_1(1+M+\cdots+M^k)$-shadows $\{y_n\}$.*

Proof. Note first that

$$\begin{aligned}
|y_{nk+1}-f_{nk}(x_{nk})|&\leqq|y_{nk+1}-f_{nk}(y_{nk})|+|f_{nk}(y_{nk})-f_{nk}(x_{nk})|\\
&\leqq\delta+M\varepsilon\leqq\varepsilon_1(1+M).
\end{aligned}$$

Then we show by induction, as in the proof of Lemma 5.2, that

$$|y_{nk+i}-(f_{nk+i-1}\circ\cdots\circ f_{nk})(x_{nk})|\leqq\varepsilon_1(1+M+\cdots+M^{i})$$

for $1\leqq i\leqq k$.

Proof of the Shadowing Lemma. Let k be a positive integer such that $16K^3\lambda^k\leqq 1$. We first prove the Shadowing Lemma for the sequence of maps $F_n=f_{nk+k-1}\circ\cdots\circ f_{nk}$ and hyperbolic sets $\{S_{nk}\}$. If we define $\omega(\eta)=\sup\{|DF_n(y)-DF_n(x)|\colon x\in S_{nk},\ y\in X_{nk},\ |y-x|\leqq\eta,\ n\in\mathbf{Z}\}$, it follows from the uniform continuity and boundedness of Df_n that $\omega(\eta)\to 0$ as $\eta\to 0$. Then we choose $\varepsilon_0>0$ so that $\varepsilon_0<\Delta$ and $\omega(\varepsilon_0)\leqq 1/(4K+2)$. Also if we define $\bar{\omega}(\eta)=\sup\{\|\mathbf{P}_n(y)-\mathbf{P}_n(x)\|\colon x\in S_n,\ y\in X_n,\ |y-x|\leqq\eta,\ n\in\mathbf{Z}\}$ it follows from the uniform continuity of $\mathbf{P}_n(x)$ that $\bar{\omega}(\eta)\to 0$ as $\eta\to 0$. Then given $0<\bar{\varepsilon}\leqq\varepsilon_0$, we let $\delta_1>0$ be such that $\delta_1\leqq\Delta$, $(4K+2)\delta_1\leqq\bar{\varepsilon}$, $8M^kK\omega(\delta_1)\leqq 1$, $4K\omega(\delta_1)\leqq 1$. (Note: M is defined before Lemma 5.2.)

Now suppose $\{\bar{y}_n\}_{n\in\mathbf{Z}}$ is a δ_1-pseudo-orbit for F_n with $\bar{y}_n\in S_{nk}$ for all n. We show the existence of a unique orbit of F_n that $\bar{\varepsilon}$-shadows $\{\bar{y}_n\}$. First we apply Lemma 3.2 to $A_n=DF_n(\bar{y}_n)$, $Q_n=\mathbf{P}_{nk}(\bar{y}_n)$. For all n, $|A_n|\leqq M^k$, $|Q_n|\leqq K$, $|I-Q_n|\leqq K$. Also by the hyperbolicity, $|A_nQ_n|\leqq K\lambda^k$ for all n and $A_n(I-Q_n)\colon\mathcal{N}(Q_n)\to\mathcal{N}(\mathbf{P}_{(n+1)k}(F_n(\bar{y}_n)))$ is invertible with inverse having norm bounded by $K\lambda^k$. Using the invariance property of $\mathbf{P}_n$,

$$\begin{aligned}
|Q_{n+1}A_n(I-Q_n)|&=|[\mathbf{P}_{(n+1)k}(\bar{y}_{n+1})-\mathbf{P}_{(n+1)k}(F_n(\bar{y}_n))]A_n(I-Q_n)|\\
&\leqq\omega(\delta_1)\cdot M^kK\leqq 1/8
\end{aligned}$$

and, similarly, $|(I-Q_{n+1})A_nQ_n|\leqq 1/8$. Also since $2K\omega(\delta_1)\leqq 1/2$ it follows from Lemma 5.1 that

$$J_n=Q_{n+1}\mathbf{P}_{(n+1)k}(F_n(\bar{y}_n))+(I-Q_{n+1})(I-\mathbf{P}_{(n+1)k}(F_n(\bar{y}_n)))$$

is invertible with $|J_n^{-1}|\leqq(1-2K\omega(\delta_1))^{-1}\leqq 2$ and that $J_n(\mathcal{N}(\mathbf{P}_{(n+1)k}(F_n(\bar{y}_n))))=\mathcal{N}(Q_{n+1})$. Hence

$$(I-Q_{n+1})A_n(I-Q_n)=J_nA_n(I-Q_n)\colon\mathcal{N}(Q_n)\to\mathcal{N}(Q_{n+1})$$

is invertible with inverse B_n satisfying $|B_n| \leqq 2K\lambda^k$ so that $|B_n(I-Q_n)| \leqq 2K^2\lambda^k$. Thus the conditions of Lemma 3.2 are satisfied with X_{nk} instead of X_n and $2K^2\lambda^k$ instead of λ. So if we define $L: \prod_{n=-\infty}^{\infty} X_{nk} \to \prod_{n=-\infty}^{\infty} X_{nk}$ by

$$(L\mathbf{u})_n = u_n - A_{n-1}u_{n-1} = u_n - DF_{n-1}(\bar{y}_{n-1})u_{n-1},$$

L is invertible with $\|L^{-1}\| \leqq 2K+1$.

Let U be the open set in $\prod_{n=-\infty}^{\infty} X_{nk}$ consisting of those $\{x_n\}$, $x_n \in X_{nk}$ satisfying $\sup_n |x_n - \bar{y}_n| < \Delta$. Then we define $\mathscr{F}: U \to \prod_{n=-\infty}^{\infty} X_{nk}$ by

$$(\mathscr{F}(\mathbf{x}))_n = x_n - F_{n-1}(x_{n-1}).$$

$\mathscr{F}$ is C^1 with $(D\mathscr{F}(\mathbf{x})\mathbf{h})_n = h_n - DF_{n-1}(x_{n-1})h_{n-1}$. $L = D\mathscr{F}(\bar{\mathbf{y}})(\bar{\mathbf{y}} = \{\bar{y}_n\})$ is invertible with $\|L^{-1}\| \leqq 2K+1$. Condition (4.1) is satisfied by choice of ε_0 and since $\|\mathscr{F}(\bar{y})\| \leqq \delta_1 \leqq \bar{\varepsilon}/(4K+2)$, condition (4.2) is also satisfied with $\bar{\varepsilon}$ instead of ε. Then it follows from Proposition 4.1 that there exists a unique $\bar{\mathbf{x}} = \{\bar{x}_n\}$ in $\prod_{n=-\infty}^{\infty} X_{nk}$ such that $\mathscr{F}(\mathbf{x}) = 0$ and $\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\| \leqq \bar{\varepsilon}$. That is, $\{\bar{x}_n\}$ is the unique orbit of $\{F_n\}$ such that $|\bar{x}_n - \bar{y}_n| \leqq \bar{\varepsilon}$ for all n.

Now let $0 < \varepsilon \leqq \varepsilon_0$ and let δ_1 correspond to $\bar{\varepsilon} = (1+M+\cdots+M^k)^{-1}\varepsilon$. Set $\delta = (1+M+\cdots+M^k)^{-1}\delta_1$ and let $\{y_n\}$ be a δ-pseudo-orbit for $\{f_n\}$ with $y_n \in S_n$ for all n. Then $\{\bar{y}_n\}_{n\in\mathbf{Z}}$, with $\bar{y}_n = y_{nk}$, is a δ_1-pseudo-orbit for $\{F_n\}$ by virtue of Lemma 5.2. So there exists a unique orbit $\{\bar{x}_n\}$ of $\{F_n\}$ which $\bar{\varepsilon}$-shadows $\{\bar{y}_n\}$. Then we define $x_{nk} = \bar{x}_n$, $n \in \mathbf{Z}$, and each x_{nk+i}, $1 \leqq i \leqq k-1$, $n \in \mathbf{Z}$, in the obvious way so that $\{x_n\}_{n\in\mathbf{Z}}$ is an orbit for $\{f_n\}$. It follows from Lemma 5.3 and the fact that $\max\{\bar{\varepsilon}, \delta\} = \bar{\varepsilon}$ that $\{x_n\}\varepsilon$-shadows $\{y_n\}$. $\{x_n\}$ must be unique because $\{x_{nk}\}$ ε-shadows $\{y_{nk}\}$ and there is a unique such orbit because $\varepsilon \leqq \varepsilon_0$.

6. Application to parabolic evolution equations. Consider the following parabolic evolution equation

$$\dot{x} + Ax = f(x, t) \tag{6.1}$$

in a Banach space X with norm $|\cdot|$. Suppose A is a sectorial operator in X (see Henry [8] for general reference in this section) with $\operatorname{Re}\sigma(A) > 0$. We can define the fractional powers $A^\alpha: \mathscr{D}(A^\alpha) \to X$, $0 \leqq \alpha \leqq 1$, and then $X^\alpha = \mathscr{D}(A^\alpha)$, the domain of A^α, becomes a Banach space with the graph norm $|x|_\alpha = |A^\alpha x|$.

We also assume that $f \in C^1(X^\alpha \times \mathbf{R}, X)$ and that f and $D_x f: X^\alpha \times \mathbf{R} \to \mathscr{L}(X^\alpha, X)$ are Lipschitzian in x and locally Hölder continuous in t. Under these conditions the initial value problem

$$\dot{x} + Ax = f(x, t), \qquad x(t_0) = x_0$$

has for all $(x_0, t_0) \in X^\alpha \times \mathbf{R}$ a unique solution

$$x(t; x_0, t_0) \in C^0([t_0, T), X^\alpha) \cap C^1((t_0, T), X) \cap C^0((t_0, T), \mathscr{D}(A)),$$

where $[t_0, T)$ is the maximal interval of existence. We denote the solution map of (6.1) by $\mathbf{T}(t, t_0)(x_0) = x(t; x_0, t_0)$.

Let $S \subset X^\alpha \times \mathbf{R}$ be a *forward invariant set* for (6.1), that is, if $(x_0, t_0) \in S$ then $\mathbf{T}(t, t_0)(x_0)$ is defined for all $t \geqq t_0$ and $(\mathbf{T}(t, t_0)(x_0), t) \in S$. This means that $\mathbf{T}(t, t_0)S_{t_0} \subset S_t$ for all $t \geqq t_0$, where $S_t = \{x \in X^\alpha : (x, t) \in S\}$ is the t-section of S. We say S is *hyperbolic* if:

(i) For $x \in S_t$, $t \in \mathbf{R}$, there is a splitting

$$X^\alpha = E_t^s(x) \oplus E_t^u(x) \tag{6.2}$$

which is *invariant*, that is,

$$\begin{aligned} D_x\mathbf{T}(t, t_0)(x)E_{t_0}^s(x) &\subset E_t^s(\mathbf{T}(t, t_0)(x)), \\ D_x\mathbf{T}(t, t_0)(x)E_{t_0}^u(x) &\subset E_t^u(\mathbf{T}(t, t_0)(x)) \end{aligned} \tag{6.3}$$

for all $x \in S_{t_0}$ and $t, t_0 \in \mathbf{R}$ with $t \geqq t_0$, and also *continuous*, that is, if $\mathbf{P}_t(x)$ is the projection with range $E_t^s(x)$ and nullspace $E_t^u(x)$, $\mathbf{P}_t(x)$ is continuous uniformly with respect to $x \in S_t$, $t \in \mathbf{R}$. We also assume that $D_x\mathbf{T}(t, t_0)(x): E_{t_0}^u(x) \to E_t^u(\mathbf{T}(t, t_0)(x))$ is an isomorphism with (bounded) inverse

$$(D_x\mathbf{T}(t, t_0)(x))^{-1}: E_t^u(\mathbf{T}(t, t_0)(x)) \to E_{t_0}^u(x).$$

(ii) There exist constants $K \geqq 1$, $\beta > 0$ such that for $x \in S_{t_0}$ and $t, t_0 \in \mathbf{R}$ with $t \geqq t_0$,

$$\begin{aligned} &|D_x\mathbf{T}(t, t_0)(x)\mathbf{P}_{t_0}(x)|_{\mathscr{L}(X^\alpha, X^\alpha)} \leqq K e^{-\beta(t-t_0)}, \\ &|(D_x\mathbf{T}(t, t_0)(x))^{-1}(I - \mathbf{P}_t(\mathbf{T}(t, t_0)(x)))|_{\mathscr{L}(X^\alpha, X^\alpha)} \leqq K e^{-\beta(t-t_0)}. \end{aligned} \tag{6.4}$$

Now we want to define pseudosolutions of (6.1). Let $\bigcup_{n \in \mathbf{Z}} [\tau_{n-1}, \tau_n] = \mathbf{R}$ be a partition of $\mathbf{R}$ with $\inf\{\tau_n - \tau_{n-1}: n \in \mathbf{Z}\} = \tau > 0$. Then, if δ is positive, we say the sequence $\{x_n(t)\}$, $t \in [\tau_{n-1}, \tau_n]$, $n \in \mathbf{Z}$ is a *δ-pseudosolution* of (6.1) if for all n

$$x_n(\cdot) \in C^0([\tau_{n-1}, \tau_n], X^\alpha) \cap C^1((\tau_{n-1}, \tau_n), X) \cap C^0((\tau_{n-1}, \tau_n), \mathscr{D}(A))$$

and

$$\sup\{|h_n(t)|: \tau_{n-1} \leqq t \leqq \tau_n\} \leqq \delta, \qquad |g_n|_\alpha \leqq \delta, \tag{6.5}$$

where $h_n \in C^0([\tau_{n-1}, \tau_n], X)$, defined by

$$h_n(t) = \dot{x}_n(t) + Ax_n(t) - f(x_n(t), t) \tag{6.6}$$

is the *residual error* and

$$g_n = x_n(\tau_n) - x_{n+1}(\tau_n) \tag{6.7}$$

is the *jump* at τ_n.

If ε is positive, a solution $x(t)$ of (6.1) is said to *ε-shadow* the δ-pseudosolution $\{x_n(t)\}$ if $x(t)$ is defined for all t and $|x(t) - x_n(t)| \leqq \varepsilon$ for $\tau_{n-1} \leqq t \leqq \tau_n$, $n \in \mathbf{Z}$.

THEOREM 6.1. *Let A, X, X^α, $f(t, x)$ be as above and suppose $S \subset X^\alpha \times \mathbf{R}$ is a forward invariant hyperbolic set for* (6.1) *such that $f(x, t)$ and $D_x f(x, t)$ are bounded and Lipschitz continuous in a Δ-neighborhood O of S in $X^\alpha \times \mathbf{R}$.*

Let $\{x_n(t)\}$, $\tau_{n-1} \leqq t \leqq \tau_n$, $n \in \mathbf{Z}$, be a δ-pseudosolution of (6.1) *such that for $\tau_{n-1} \leqq t \leqq \tau_n$ and $n \in \mathbf{Z}$, $x_n(t)$ is in a δ-neighborhood of S_t in the X^α norm.*

Then there exist $\varepsilon_0 > 0$ and a positive function $\delta(\varepsilon)$, both depending only on A, f, $\tau = \inf_n(\tau_n - \tau_{n-1})$, such that if $0 < \varepsilon \leqq \varepsilon_0$ and $\delta \leqq \delta(\varepsilon)$, there is a unique solution of (6.1) *that ε-shadows $\{x_n(t)\}$.*

For the proof of Theorem 6.1, we need a lemma.

LEMMA 6.2. *Let the hypotheses of Theorem* 6.1 *hold and let M be the bound for $\mathscr{D}_x f(t, x)$ in O. Let $x(t)$ be a solution of* (6.1) *in S and let $y(t)$ be a solution of the initial value problem*

$$\dot{y} + Ay = f(y, t) + h(t), \qquad y(t_0) = \xi \tag{6.8}$$

for $t \in [t_0, t_0 + 2\tau]$, where $|\xi - x(t_0)|_\alpha \leqq \delta$ and $h \in C^0([t_0, t_0 + 2\tau], X)$ with $\sup|h(t)| \leqq \delta$. If $(y(t), t) \in O$ for $t \in [t_0, t_0 + 2\tau]$, then there exists a constant $C \geqq 1$ depending only on A and M such that $|y(t) - x(t)|_\alpha \leqq C\delta$ on $[t_0, t_0 + 2\tau]$.

Proof. Our assumptions on A imply the existence of positive constants C_1, C_2 and a such that for $t \geqq 0$

$$|e^{-At}|_{\mathscr{L}(X^\alpha, X^\alpha)} \leqq C_1 e^{at}, \qquad |e^{-At}|_{\mathscr{L}(X, X^\alpha)} \leqq C_2 t^{-\alpha} e^{at}.$$

Now $z(t)=y(t)-x(t)$ satisfies the integral equation

$$z(t)=e^{-A(t-t_0)}z(t_0)+\int_{t_0}^{t}e^{-A(t-s)}\{f(x(s)+z(s),s)-f(x(s),s)+h(s)\}\,ds.$$

Then for $t_0\leqq t\leqq t_0+2\tau$,

$$|z(t)|_\alpha\leqq C_1\,e^{a(t-t_0)}\delta+\int_{t_0}^{t}MC_2(t-s)^{-\alpha}\,e^{a(t-s)}|z(s)|_\alpha\,ds+\int_{t_0}^{t}C_2(t-s)^{-\alpha}\,e^{a(t-s)}\delta\,ds.$$

It follows from an inequality in Henry [8, Lemma 7.1.1, p. 188] that $|z(t)|_\alpha<C\delta$ for $t_0\leqq t\leqq t_0+2\tau$. The proof is completed.

Proof of Theorem 6.1. The hypotheses on A and f, Lemma 6.2, and Henry [8] imply that there exists a closed Δ_1-neighborhood O_1 of S in $X^\alpha\times\mathbf{R}$ such that for $(x,t_0)\in O_1$, $\mathbf{T}(t,t_0)(x)$ is defined for $t_0\leqq t\leqq t_0+2\tau$ and both $\mathbf{T}(t,t_0)(x)$ and $D_x\mathbf{T}(t,t_0)(x)$ are bounded and continuous, uniformly with respect to $(x,t_0)\in O_1$ and $t\in[t_0,t_0+2\tau]$. (These functions have ranges in X^α and $\mathscr{L}(X^\alpha,X^\alpha)$, respectively, and the continuity is with respect to these norms.)

Without loss of generality we may assume that $0\leqq\tau\leqq\tau_n-\tau_{n-1}\leqq 2\tau$ for all n. We first consider the case where $h_n(t)=0$ and $x_n(t)\in S_t$ for all t and n. Then if we let X_n be X^α for all n and f_n be $\mathbf{T}(\tau_n,\tau_{n-1}):X^\alpha\to X^\alpha$ the domain of f_n contains a closed Δ_1-neighborhood of $S_{\tau_{n-1}}$ in which f_n and Df_n are both bounded and uniformly continuous, uniformly with respect to $n\in\mathbf{Z}$. From the hyperbolicity of S with respect to (6.1), we see that $\{S_{\tau_{n-1}}\}_{n\in\mathbf{Z}}$ is invariant ($f_n(S_{\tau_{n-1}})\subset S_{\tau_n}$) and hyperbolic for $\{f_n\}_{n\in\mathbf{Z}}$ with projections $\mathbf{P}_{\tau_{n-1}}(x)$ and constants K, $e^{-\beta\tau}$. Hence conditions (i), (ii), (iii) of the Shadowing Lemma hold. Set $y_n=x_n(\tau_{n-1})$, $n\in\mathbf{Z}$. Then $y_n\in S_{\tau_{n-1}}$ for all n and

$$|f_n(y_n)-y_{n+1}|_\alpha=|x_n(\tau_n)-x_{n+1}(\tau_n)|_\alpha\leqq\delta.$$

So if $0<\varepsilon\leqq\varepsilon_1$ and $\delta\leqq\delta_1(\varepsilon)$ (ε_1 and $\delta_1(\varepsilon)$ correspond to ε_0 and $\delta(\varepsilon)$ in the Shadowing Lemma) there is a unique solution $x(t)$ of (6.1) such that $|x(\tau_{n-1})-x_n(\tau_{n-1})|_\alpha\leqq\varepsilon$ for all n.

Now we consider the general case. We suppose $0<\varepsilon\leqq\varepsilon_0=\frac{1}{2}\min\{\Delta,\varepsilon_1\}$ and $\delta\leqq\delta(\varepsilon)=\min\{(2C+1)^{-1}\delta_1(\varepsilon/2C),\varepsilon/2C\}$. Let $\{x_n(\cdot)\}$ be a δ-pseudosolution as in the statement of the theorem. Since for all n, $x_n(\tau_{n-1})$ is in a δ-neighborhood of $S_{\tau_{n-1}}$, we can choose y_n in $S_{\tau_{n-1}}$ so that $|x_n(\tau_{n-1})-y_n|_\alpha\leqq\delta$. Then let $\bar{x}_n(t)$ be the solution of (6.1) satisfying $\bar{x}_n(\tau_{n-1})=y_n$. By Lemma 6.2 with $t_0=\tau_{n-1}$, $x(t)=\bar{x}_n(t)$, $h(t)=h_n(t)$, $y(t)=x_n(t)$ we have

$$|\bar{x}_n(t)-x_n(t)|_\alpha\leqq C\delta$$

for $\tau_{n-1}\leqq t\leqq\tau_n$. This holds for $n\in\mathbf{Z}$. Moreover,

$$\begin{aligned}|\bar{x}_n(\tau_n)-\bar{x}_{n+1}(\tau_n)|_\alpha&\leqq|\bar{x}_n(\tau_n)-x_n(\tau_n)|_\alpha+|x_n(\tau_n)-x_{n+1}(\tau_n)|_\alpha+|x_{n+1}(\tau_n)-\bar{x}_{n+1}(\tau_n)|_\alpha\\&\leqq(2C+1)\delta\leqq\delta_1(\varepsilon/2C).\end{aligned}\tag{6.9}$$

Hence $\{\bar{x}_n(t)\}$ is a $\delta_1(\varepsilon/2C)$-pseudosolution of (6.1) with $\bar{x}_n(t)\in S_t$ for all t and n, where $\varepsilon/2C<\varepsilon_1$. It follows from the first part of the proof that there is a unique solution $x(t)$ of (6.1) such that

$$|x(\tau_{n-1})-\bar{x}_n(\tau_{n-1})|_\alpha\leqq\varepsilon/2C\tag{6.10}$$

for all n. Then for all n

$$\begin{aligned}|x(\tau_{n-1})-x_n(\tau_{n-1})|_\alpha&\leqq|x(\tau_{n-1})-\bar{x}_n(\tau_{n-1})|_\alpha+|\bar{x}_n(\tau_{n-1})-x_n(\tau_{n-1})|_\alpha\\&\leqq\varepsilon/2C+\delta.\end{aligned}$$

By Lemma 6.2 with $t_0 = \tau_{n-1}$, $x(t) = \bar{x}_n(t)$, $y(t) = x(t)$ and $\varepsilon/2C + \delta$ instead of δ (note: $C(\varepsilon/2C + \delta) = \varepsilon/2 + C\delta < \Delta$), we deduce for $\tau_{n-1} \leqq t \leqq \tau_n$, $n \in \mathbf{Z}$ that

$$|x(t) - x_n(t)|_\alpha \leqq \varepsilon/2 + C\delta \leqq \varepsilon.$$

That is, $x(t)$ does ε-shadow the δ-pseudosolution $\{x_n(t)\}$.

Let $\tilde{x}(t)$ be another such solution. Then for all n

$$\begin{aligned}|\tilde{x}(\tau_{n-1}) - \bar{x}_n(\tau_{n-1})|_\alpha &\leqq |\tilde{x}(\tau_{n-1}) - x_n(\tau_{n-1})|_\alpha + |x_n(\tau_{n-1}) - \bar{x}_n(\tau_{n-1})|_\alpha \\ &\leqq \varepsilon + \delta \leqq 3\varepsilon/2 < \varepsilon_1,\end{aligned}$$

where for all n, using (6.9),

$$\begin{aligned}|f_n(\bar{x}_n(\tau_{n-1})) - \bar{x}_{n+1}(\tau_n)|_\alpha &= |\bar{x}_n(\tau_n) - \bar{x}_{n+1}(\tau_n)|_\alpha \\ &\leqq \delta_1(\varepsilon/2C) \\ &\leqq \delta_1(3\varepsilon/2).\end{aligned}$$

(Note: we assume without loss of generality that $\delta_1(\varepsilon)$ is nondecreasing in ε.) That is, the sequence $\{\tilde{x}(\tau_{n-1})\}$ is an orbit of $\{f_n\}$ that $3\varepsilon/2$-shadows the $\delta_1(3\varepsilon/2)$-pseudo-orbit $\{\bar{x}_n(\tau_{n-1})\}$, where $3\varepsilon/2 < \varepsilon_1$. But by (6.10), $\{x(\tau_{n-1})\}$ is another such sequence and so it follows by uniqueness that $\tilde{x}(\tau_{n-1}) = x(\tau_{n-1})$ for all n. Hence $x(t)$ is unique.

REFERENCES

[1] D. V. ANOSOV, *Geodesic flows on compact Riemannian manifolds of negative curvature*, Proc. Steklov Inst. Math., 90 (1967); Amer. Math. Soc. Trans., 1969.

[2] C. M. BLAZQUEZ, *Transverse homoclinic orbits in periodically perturbed parabolic equations*, Nonlinear Anal., 10 (1986), pp. 1277-1291.

[3] R. BOWEN, *ω-limit sets for Axiom A diffeomorphisms*, J. Differential Equations, 18 (1975), pp. 333-339.

[4] C. C. CONLEY, *Hyperbolic invariant sets and shift automorphisms*, in Dynamical Systems Theory and Applications, J. Moser, ed., Lecture Notes in Physics 38, Springer-Verlag, Berlin, New York, 1975, pp. 539-549.

[5] I. EKELAND, *Some lemmas about dynamical systems*, Math. Scand., 52 (1983), pp. 262-268.

[6] J. GUCKENHEIMER, J. MOSER, AND S. NEWHOUSE, *Dynamical Systems*, Birkhäuser, Boston, 1980.

[7] S. M. HAMMEL, J. A. YORKE, AND C. GREBOGI, *Do numerical orbits of chaotic dynamical processes represent true orbits*, J. Complexity, 3 (1987), pp. 136-145.

[8] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, 1981.

[9] U. KIRCHGRABER, *Erratische Lösungen der periodisch gestörten Pendulgleichung*, preprint, University of Würzburg, 1982.

[10] O. E. LANFORD III, *Introduction to the mathematical theory of dynamical systems*, in Chaotic Behavior of Deterministic Systems, Les Houches, 1981, North-Holland, Amsterdam, New York, 1983, pp. 3-51.

[11] X. B. LIN, *Shadowing lemma and singularly perturbed boundary value problems*, SIAM J. Appl. Math., submitted.

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

[13] ———, *A perturbation theorem for exponential dichotomies*, Proc. Roy. Soc. Edinburgh, 106 (A) (1987), pp. 25-37.

[14] ———, *Exponential dichotomies, the shadowing lemma and transversal homoclinic points*, Dynamics Reported, 1 (1988), pp. 265-306.

[15] C. ROBINSON, *Stability theorems and hyperbolicity in dynamical systems*, Rocky Mountain J. Math., 7 (1977), pp. 425-437.

[16] M. SHUB, *Global Stability of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1987.

[17] D. STOFFER, personal communication.

[18] H.-O. WALTHER, *A shadowing lemma for nonreversible maps in infinite dimensional spaces*, preprint.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 558-566, May 1989

004

QUENCHING FOR SEMILINEAR SINGULAR PARABOLIC PROBLEMS*

C. Y. CHAN† AND HANS G. KAPER‡

Abstract. Let f be a real-valued function that is nondecreasing and continuously differentiable on $[0, c)$ for some finite c ($c>0$), satisfying the conditions $f(0)>0$ and $\lim_{u\to c} f(u)=\infty$. This article is concerned with positive solutions u of the semilinear singular parabolic differential equation $u_t = u_{xx}+(b/x)u_x+f(u)$, $b<1$, on a bounded interval $(0, a)$, which satisfy the initial condition $u(x, 0)=0$ and the boundary conditions $u(0, t)=0$ and $u_x(a, t)=0$. Let $\|\cdot\|$ denote the sup-norm over the interval $[0, a]$. It is shown that a solution u quenches (i.e., there exists a $T<\infty$ such that $\lim_{t\to T, t<T}\|u_t(\cdot, t)\|=\infty$) if $\|u(\cdot, t)\|$ tends to c from below as t approaches T. Furthermore, there exists a critical length a^* such that u may exist for all $t>0$ if $a<a^*$, but $\|u(\cdot, t)\|$ tends to c in finite time if $a>a^*$. A numerical method is given to compute a^*. An upper bound for the quenching time T is obtained. An example is given to illustrate the results.

Key words. semilinear singular parabolic equation, nonlinear heat equation, quenching, quenching time, critical length, numerical method

AMS(MOS) subject classifications. 35K20, 35K55

1. Introduction. The concept of *quenching* of the solution of a nonlinear heat equation was first introduced by Kawarada [8], who studied the following problem:

$$u_t = u_{xx}+(1-u)^{-1}, \qquad (x, t)\in(0, l)\times(0, T), \tag{1.1a}$$

$$u(x, 0)=0, \quad x\in(0, l), \quad u(0, t)=0, \quad u(l, t)=0, \quad t\in(0, T). \tag{1.1b}$$

The solution u of (1.1a, b) *quenches* if there exists a $T<\infty$ such that

$$\lim_{t\to T, t<T}\|u_t(\cdot, t)\|=\infty. \tag{1.2}$$

Here, $\|\cdot\|$ denotes the sup-norm over the interval $[0, l]$. The value of T is referred to as the *quenching time.*

If the solution u of (1.1a, b) quenches at some finite time T, then

$$\lim_{t\to T, t<T}\|u(\cdot, t)\|=1. \tag{1.3}$$

Kawarada has claimed that (1.3) implies quenching of the solution of (1.1a, b). If this claim were correct, it would follow that the conditions (1.2) and (1.3) are equivalent and either can be taken to define quenching. However, Kawarada has assumed without justification that a function $\hat{W}$, constructed in the course of the proof, satisfied the heat equation on curves $s^{(i)}(t)$ for t arbitrarily close to the quenching time T. Hence, Kawarada's claim needs to be reexamined. Furthermore, it would be desirable to extend Kawarada's claim, if true, to nonlinear heat equations with a general forcing term f. Since Kawarada has made use of the explicit expression $f(u)=(1-u)^{-1}$, the extension of his proof is not obvious.

* Received by the editors November 4, 1987; accepted for publication (in revised form) July 5, 1988.

† Department of Mathematics, University of Southwestern Louisiana, Lafayette, Louisiana 70504-1010. The work of this author was 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 during the summer of 1987, while the author was a Visiting Scientist in the Mathematics and Computer Science Division of Argonne National Laboratory, Argonne, Illinois, and by the Board of Regents of the State of Louisiana under grant LEQSF(86-89)-RD-A-11.

‡ Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439-4844. The work of this author was 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.

In his investigation [14] of (1.1a, b), Walter has adopted (1.3) as the definition of quenching. The same definition has been adopted by Acker and Walter [1], [2], who have studied the solutions of the general equations

$$u_t = u_{xx} + f(u), \qquad (x, t) \in (0, l) \times (0, T), \tag{1.4}$$

$$u_t = u_{xx} + f(u, u_x), \qquad (x, t) \in (0, l) \times (0, T)$$

subject to the initial and boundary conditions (1.1b). The results of [1] and [2] imply that there is a critical length l^* such that a solution u exists globally (i.e., for all $t > 0$) if $l < l^*$, but u quenches if $l > l^*$. Results on the behavior of the solution of (1.4) and (1.1b) at $l = l^*$ were given subsequently by Levine and Montgomery [11].

An upper bound for l^* for the problem (1.1a, b) can be inferred from Kawarada's article [8], viz., $l^* \leqq 2\sqrt{2} = 2.8284$. Walter [14] has shown that the critical length l^* for (1.1a, b) is in the range $1.5303 \leqq l^* \leqq \frac{1}{2}\pi = 1.5708$. The exact value of l^* has been identified by Acker and Walter [1] as $2M\sqrt{2}$, where M is the maximum value of Dawson's integral, $F(x) = e^{-x^2}\int_0^x e^{t^2}\,dt$, on the interval $(0, \infty)$ (cf. [15, § 7.1]). The numerical value of l^* is 1.53030416 (to eight decimal places).

Further results and references on the phenomenon of quenching for solutions of nonlinear equations can be found in the survey article by Levine [10].

In this article we will study the semilinear singular problem

$$u_t = u_{xx} + (b/x)u_x + f(u), \qquad (x, t) \in (0, a) \times (0, T), \tag{1.5a}$$

$$u(x, 0) = 0, \quad x \in (0, a), \quad u(0, t) = 0, \quad u_x(a, t) = 0, \quad t \in (0, T). \tag{1.5b}$$

Here f is a real-valued function that is nondecreasing and continuously differentiable on the interval $[0, c)$ for some $c > 0$, satisfying the conditions $f(0) > 0$ and $\lim_{u \to c} f(u) = \infty$. The constant b satisfies the inequality $b < 1$. If $b = 0$, then (1.5a, b) reduces to (1.1a, b) with $l = 2a$, because the solution of (1.1a, b) is symmetric about $x = \frac{1}{2}l$. We remark that if $f(0) = 0$, then $u = 0$ is the only solution of (1.5).

Throughout this article we use the abbreviation $\Omega = (0, a) \times (0, T)$. We always assume that T is maximal. We use the symbol L to denote the differential expression $Lu = u_{xx} + (b/x)u_x - u_t$.

The transformation $u(x, t) = v(z, t)$, where $z = \frac{1}{4}x^2$, reduces the degenerate elliptic-parabolic expression $zv_{zz} + \frac{1}{2}(1 + b)v_z - v_t$ to Lu. This degenerate expression arises in probability theory; it has been studied by Brezis, Rosenkrantz, Singer, and Lax [6] for $b > -1$. Also, the stochastic process described by the expression $\frac{1}{2}v_{zz} + \frac{1}{2}(b/z)v_z - v_t$, which has been studied by Lamperti [9] for $b > -1$, reduces to Lu on the transformation $u(x, t) = v(z, t)$, where $z = 2^{-1/2}x$.

The existence of unique solutions of nonhomogeneous problems for the linear differential expression L has been studied by Alexiades [3], [4] and by Alexiades and Chan [5]. In particular, it follows from Alexiades [4] that initial boundary value problems described by equations of the form $Lu = g(x, t)$, where g is a given function on Ω, and the initial and boundary conditions (1.5b), have unique classical solutions if $b < 1$.

In § 2, we prove that the (unique) solution u of (1.5a, b) quenches at some finite time T if $\lim_{t \to T, t < T} \|u(\cdot, t)\| = c$. This result proves and generalizes Kawarada's claim; our method of proof is, however, different from Kawarada's. In § 3, we establish the existence of a critical length a^* and show that a^* is determined by the solution of the corresponding steady state problem. Picard's iteration scheme gives a strictly monotone sequence of functions, which converges upwards to the minimal steady state solution. We also give a procedure to compute a^*. In § 4, we obtain an upper bound for the

quenching time T by solving a singular Sturm-Liouville problem and using a comparison result. In § 5, we illustrate our results by considering the special case $f(u) = (1-u)^{-1}$.

2. Quenching. In this section we prove that (1.5a, b) has at most one solution. This solution is necessarily positive (i.e., nonnegative and nontrivial) and nondecreasing in each variable separately. Assuming that the solution exists before quenching time, we show that it quenches at a finite time T if the value of u at a approaches c, the critical value of f, as t tends to T from below.

LEMMA 1. *The initial boundary value problem* (1.5a, b) *has at most one solution u. This solution has the following properties*: (i) $u(x, t) > 0$ *for all* $(x, t) \in \Omega \cup (\{a\} \times (0, T))$; (ii) $u(x, t)$ *is a strictly increasing function of* t *for each* $x \in (0, a]$; (iii) $u(x, t)$ *is a nondecreasing function of* x *for each* $t \in (0, T)$.

Proof. Let u_1 and u_2 be two distinct solutions of (1.5a, b) and let $y = u_1 - u_2$. Then y satisfies the differential equation $[L + f'(\eta)]y = 0$ in Ω for some η between u_1 and u_2, the initial condition $y(x, 0) = 0$ and the boundary conditions $y(0, t) = 0$ and $y_x(a, t) = 0$. Since $f'(\eta)$ is bounded above, the uniqueness of u follows from the strong maximum principle and the parabolic version of Hopf's lemma (cf. [13, §§ 3.2, 3.3]).

(i) Because $f(0) > 0$, we have $Lu + f(u) - f(0) < 0$ in Ω, so $[L + f'(\eta)]u < 0$ for some η between 0 and u. It follows from the strong maximum principle and the parabolic version of Hopf's lemma that $u > 0$ in $\Omega \cup (\{a\} \times (0, T))$.

(ii) For any $h \in (0, T)$, let u_h be defined in $\Omega_h = (0, a) \times (0, T-h)$ by the expression $u_h(x, t) = u(x, t+h)$ and let $y = u_h - u$. Then $[L + f'(\zeta)]y = 0$ in Ω_h for some ζ between u and u_h. On the parabolic boundary $\partial_p \Omega_h$ of Ω_h we have $y(x, 0) > 0$, $y(0, t) = 0$, and $y_x(a, t) = 0$. The inequality $y > 0$ in Ω_h follows from the strong maximum principle and the parabolic version of Hopf's lemma; hence, $u(x, t)$ is a strictly increasing function of t for each $x \in (0, a)$.

(iii) For any $\varepsilon \in (0, a)$, let $\Omega_\varepsilon = (\varepsilon, a) \times (0, T)$. As in the proof of part (i), we show that the solution u_ε of the (regular) problem

$$Lu_\varepsilon = -f(u_\varepsilon), \qquad (x, t) \in \Omega_\varepsilon, \tag{2.1a}$$

$$u_\varepsilon(x, 0) = 0, \quad u_\varepsilon(\varepsilon, t) = 0, \quad u_{\varepsilon,x}(a, t) = 0, \tag{2.1b}$$

is positive in $\Omega_\varepsilon \cup (\{a\} \times (0, T))$. (Here, $u_{\varepsilon,x}$ denotes the partial derivative of u_ε with respect to x.) Differentiating (2.1a) with respect to x, we obtain

$$[L + f'(u_\varepsilon) - b/x^2]u_{\varepsilon,x} = 0, \qquad (x, t) \in \Omega_\varepsilon.$$

It follows from (2.1b) that $u_{\varepsilon,x}(x, 0) = 0$, $u_{\varepsilon,x}(\varepsilon, t) \geqq 0$, and $u_{\varepsilon,x}(a, t) = 0$, so the strong maximum principle implies that $u_{\varepsilon,x} > 0$ in Ω_ε. It also follows from the strong maximum principle and the parabolic version of Hopf's lemma that u_ε is strictly monotone increasing as ε decreases. In particular, we have $0 < u_\varepsilon < u$ in Ω_ε. Since u_ε is bounded, $\lim_{\varepsilon \to 0} u_\varepsilon$ exists; we denote it by Z. Thus, $Z_x \geqq 0$ and $0 < u_\varepsilon \leqq Z \leqq u$ in Ω_ε.

For any $\sigma \in (\varepsilon, a)$, let $\Omega_\sigma = (\sigma, a) \times (0, T)$. We consider the solution u_σ of the (regular) problem

$$Lu_\sigma = -f(u_\sigma), \qquad (x, t) \in \Omega_\sigma,$$

$$u_\sigma(x, 0) = 0, \quad u_\sigma(\sigma, t) = u_\varepsilon(\sigma, t), \quad u_{\sigma,x}(a, t) = 0.$$

Let L^* denote the adjoint of L (defined with the appropriate adjoint boundary conditions) and let $R^*(\xi, \tau; x, t)$ be its Green's function (cf. [7, § 3.7; Chap. 5, Problem 5] and [12, § 6.2]). We take $u = u_\varepsilon$ and $v(\xi, \tau) = R^*(\xi, \tau; x, t)$ in Green's identity,

$$vLu - uL^*v = [vu_x - uv_x + (b/x)uv]_x - (uv)_t.$$

Integrating over the domain $(\sigma, a)\times(0, t-\delta)$, where δ is some small positive constant less than t, using Green's theorem, and letting δ tend to zero, we find

$$u_\varepsilon(x,t)=\int_0^t\int_\sigma^a R^*(\xi,\tau;x,t)f(u_\varepsilon(\xi,\tau))\,d\xi\,d\tau + \int_0^t R_\xi^*(\sigma,\tau;x,t)u_\varepsilon(\sigma,\tau)\,d\tau, \qquad (x,t)\in\Omega_\sigma.$$

Since $R^*(\xi,\tau;x,t)>0$ for $(\xi,\tau)\in(\sigma,a)\times(0,t)$ (cf. [7, § 3.7]), it follows that $R_\xi^*(\sigma,\tau;x,t)\geqq 0$. Because u_ε and $f(u_\varepsilon)$ are nondecreasing as ε decreases, it follows from the monotone convergence theorem that

$$Z(x,t)=\int_0^t\int_\sigma^a R^*(\xi,\tau;x,t)f(Z(\xi,\tau))\,d\xi\,d\tau + \int_0^t R_\xi^*(\sigma,\tau;x,t)Z(\sigma,\tau)\,d\tau, \qquad (x,t)\in\Omega_\sigma.$$

Thus, $LZ=-f(Z)$ in Ω_σ. Since σ is arbitrary it follows that $LZ=-f(Z)$ in Ω. Also, $Z_x(a,t)=0$ and $Z(x,0)=0$. Because $0\leqq u_\varepsilon\leqq Z\leqq u$ in Ω, it must be the case that $Z(0,t)=0$. Since u is unique, we have $u=Z$; hence, $u_x\geqq 0$. □

THEOREM 2. *Suppose that the function f is such that*

$$\int_0^c f(u)\,du=\infty. \tag{2.2}$$

If

$$\lim_{t\to T} u(a,t)=c, \tag{2.3}$$

for some finite T, then the solution u of (1.5a, b) *quenches.*

Proof. The proof is by contradiction, where we assume that (2.3) holds, but $\|u_t(\cdot,t)\|$ remains bounded as $u(a,t)$ tends to c. (We recall from Lemma 1(iii) that $\|u(\cdot,t)\|=u(a,t)$.)

By assumption, there exists a constant M such that $u_t(x,t)\leqq M$ for all $(x,t)\in\bar{\Omega}$, the closure of Ω. Hence,

$$u_{xx}+(b/x)u_x=x^{-b}(x^b u_x)_x\leqq M-f(u), \qquad (x,t)\in\Omega. \tag{2.4}$$

Because u is a nondecreasing function of each argument and $u(a,t)$ tends to c as $t\to T$, there certainly exists an open rectangle $Q=(\xi,a)\times(\tau,T)$ with $\xi>0$ such that $f(u)\geqq 2M$ in Q. Then

$$2x^{-b}(x^b u_x)_x\leqq -f(u), \qquad (x,t)\in Q. \tag{2.5}$$

If $b<0$, it follows that $2(x^b u_x)(x^b u_x)_x\leqq -f(u)u_x/a^{-2b}$. On integration from ξ to a we find

$$u_x^2(\xi,t)\geqq\left(\frac{\xi}{a}\right)^{-2b}\int_{u(\xi,t)}^{u(a,t)} f(u)\,du, \qquad t\in(\tau,T).$$

As t approaches T, the integral grows beyond bounds, because of (2.2), so the same must be true for the (nonnegative) quantity $u_x(\xi,t)$. If $b\geqq 0$, then it follows from (2.5) that $2u_{xx}\leqq -f(u)$. Since $u_x\geqq 0$, we have

$$u_x^2(\xi,t)\geqq\int_{u(\xi,t)}^{u(a,t)} f(u)\,du, \quad t\in(\tau,T).$$

Again we arrive at the conclusion that $u_x(\xi,t)$ grows beyond bounds as t tends to T.

Since f is nondecreasing and u is nonnegative, we have $f(u) > 0$. The inequality (2.4) therefore yields $(x^b u_x)_x \leqq M x^b$ in Ω. It follows on integration from some point x to ξ that

$$u_x(x, t) \geqq x^{-b}\{\xi^b u_x(\xi, t) - [M/(1+b)](\xi^{1+b} - x^{1+b})\}, \qquad b \neq -1,$$

$$u_x(x, t) \geqq (x/\xi) u_x(\xi, t) - Mx \ln(\xi/x), \qquad b = -1.$$

A second integration from zero to ξ yields

$$u(\xi, t) \geqq \frac{\xi u_x(\xi, t)}{1-b} - \frac{M\xi^2}{1+b}\left(\frac{1}{1-b} - \frac{1}{2}\right), \qquad b \neq -1,$$

$$u(\xi, t) \geqq \frac{1}{2}\xi u_x(\xi, t) - \frac{1}{4} M\xi^2, \qquad b = -1.$$

As t tends to T, these lower bounds become arbitrarily large, so it would follow that $u(\xi, t)$ becomes aribtrarily large. But now we have a contradiction, because $u(\xi, t)$ is less than $u(a, t)$ and the latter quantity tends to the finite limit c. □

3. Critical length. In this section we establish the existence of a critical length and show how the critical length can be determined.

THEOREM 3. *If $T = \infty$ and there exists a constant $C \in (0, c)$ such that $u(x, t) \leqq C$ for all $(x, t) \in \Omega$, then $u(\cdot, t)$ converges from below to a solution U of the singular nonlinear two-point boundary value problem*

$$U'' + (b/x)U' = -f(U), \quad x \in (0, a), \quad U(0) = 0, \quad U'(a) = 0. \tag{3.1}$$

The convergence is uniform on $[0, a]$. Furthermore, $u < U$ on $(0, a] \times (0, \infty)$.

Proof. Since the homogeneous problem corresponding to (3.1) has only the trivial solution, Green's function $G(x; y)$ for (3.1) exists. A direct computation gives

$$G(x; y) = \begin{cases} (1-b)^{-1} x^{1-b}, & 0 \leqq x \leqq y, \\ (1-b)^{-1} y^{1-b}, & y \leqq x \leqq a. \end{cases}$$

Let F denote the function

$$F(x, t) = \int_0^a y^b G(x; y) u(y, t)\, dy, \qquad (x, t) \in \Omega. \tag{3.2}$$

If u satisfies the initial boundary value problem (1.5a, b), then Green's identity yields

$$F_t(x, t) = -u(x, t) + \int_0^a y^b G(x; y) f(u(y, t))\, dy. \tag{3.3}$$

According to Lemma 1(ii), u is strictly increasing in t for $x \in (0, a]$. Since f is nondecreasing, the integrand in (3.3) is monotone nondecreasing with respect to t. It follows from the monotone convergence theorem and the continuity of f that

$$\lim_{t\to\infty} F_t(x, t) = -\lim_{t\to\infty} u(x, t) + \int_0^a y^b G(x; y) f(\lim_{t\to\infty} u(y, t))\, dy.$$

From (3.2) and Lemma 1(ii) we infer that $\lim_{t\to\infty} F_t(x, t) \geqq 0$. We claim that the limit is exactly zero. If the limit were (strictly) positive at some point x, it would follow that $F(x, t)$ would increase without bound as t tends to infinity, so u would reach c in a

finite time, contradictory to the assumption that T is infinite. The identity (3.3) implies that

$$\lim_{t\to\infty} u(x,t) = \int_0^a y^b G(x;y) f(\lim_{t\to\infty} u(y,t))\, dy.$$

That is, $\lim_{t\to\infty} u(x,t) = U(x)$. The uniform convergence follows from Dini's theorem.

Lemma 1(i) and (ii) imply that $U>0$ on $(0,a]$. Furthermore, $[L+f'(\eta)](U-u) = 0$ in Ω for some η between U and u; $U-u>0$ at $t=0$, $U-u=0$ at $x=0$, and $U'-u_x=0$ at $x=a$. Hence, $U-u>0$ for $0<x\leqq a$. □

THEOREM 4. *Let u_α denote the solution of the problem* (1.5a, b), *where a is replaced by $a+\alpha$ for some constant $\alpha>0$. If $\lim_{t\to\infty} u(a,t)=c$, then u_α quenches.*

Proof. Assume that u_α does not quench. Let $y=u_\alpha-u$. Then $[L+f'(\eta)]y=0$ in Ω for some η between u and u_α. Furthermore, $y(x,0)=0$, $y(0,t)=0$, and $y_x(a,t)\geqq 0$ by Lemma 1(iii). Therefore, $u_\alpha\geqq u$ in Ω.

Let ε and t_0 be positive numbers, chosen so that $f(z)\geqq(4\varepsilon/\alpha)(2/\alpha+3|b|/a)+\alpha^2$ for $z\in[c-\varepsilon, c)$ and $u(a,t_0)\geqq c-\varepsilon$. Let S denote the domain $(a,a+\alpha)\times(t_0,\infty)$.

By assumption, u_α exists for all $t>0$; in particular, it must be the case that $u_\alpha<c$ in S. It follows from Lemma 1(ii) and (iii) that $u_\alpha\geqq c-\varepsilon$ on the parabolic boundary $\partial_p S$ of S. Consider the function z, defined by the expression $z(x,t)=c-\varepsilon+(x-a)(a+\alpha-x)(t-t_0)$ on S. Clearly, $z=c-\varepsilon$ on $\partial_p S$. Furthermore,

$$Lz+f(z)\geqq -2(t-t_0)+(b/x)[2(a-x)+\alpha](t-t_0)$$
$$-(x-a)(a+\alpha-x)+(4\varepsilon/\alpha)(2/\alpha+3|b|/a)+\alpha^2, \qquad z\in[c-\varepsilon, c).$$

The lower bound is nonnegative in $(a,a+\alpha)\times(t_0, t_0+4\varepsilon/\alpha^2)$. It follows from the strong maximum principle that $u_\alpha>z$ in this domain; hence, the (nonstrict) inequality $u_\alpha\geqq z$ holds at the point $(x,t)=(a+\alpha/2, t_0+4\varepsilon/\alpha^2)$, where z assumes the value c. But now we have a contradiction, since $u_\alpha<c$ everywhere in S. □

Theorem 3 implies that there exists a critical length a^* such that u exists on $[0,a]$ for all $t>0$ if $a<a^*$. The critical length a^* is determined as the supremum of all positive values a for which a solution U of (3.1) exists; if $U(a^*)$ exists, then $u(a^*,t)$ exists also. According to Theorem 4, u quenches if $a>a^*$.

We now give a procedure to compute the critical length a^*. Let $a<a^*$ and let $U_0=0$ on $[0,a]$. We construct a sequence $\{U_n\}_{n\in N}$ by defining U_n as the solution of the boundary value problem

$$U_n''+(b/x)U_n'+f(U_{n-1})=0, \quad x\in(0,a), \quad U_n(0)=0, \quad U_n'(a)=0.$$

In terms of Green's function G, we have

$$U_n(x) = \int_0^a \xi^b G(x;\xi) f(U_{n-1}(\xi))\, d\xi, \qquad n=1,2,\cdots. \tag{3.4}$$

The sequence is well defined.

THEOREM 5. *The sequence $\{U_n\}_{n\in N}$ converges monotonically upwards to the minimal solution U of the boundary value problem* (3.1). *This minimal solution satisfies the inequality $U<c$ on $[0,a]$.*

Proof. Since $f(0)>0$ and $G(\cdot;\xi)>0$ in Ω, it follows that $U_1>0$ on $(0,a]$. From Lemma 1(ii) and Theorem 3 it follows that $0<U<c$ on $(0,a]$. Since f is nondecreasing, we have $(U-U_1)''+(b/x)(U-U_1)'\leqq 0$; furthermore, $(U-U_1)(0)=0$ and $(U-U_1)'(a)=0$. The positivity of Green's function then yields the inequality $U>U_1$ on $(0,a]$. Similarly $U>U_2>U_1$ on $(0,a]$ and, by induction,

$$0<U_n<U_{n+1}<U<c \quad \text{on } (0,a], \quad n=1,2,\cdots. \tag{3.5}$$

Hence, there exists a function $V(x)$ such that $\lim_{n\to\infty} U_n = V$. The integrand in (3.4) is nondecreasing with respect to n and integrable, so it follows from the monotone convergence theorem that V satisfies the equation

$$V(x) = \int_0^a \xi^b G(x;\xi) f(V(\xi))\, d\xi.$$

Hence, V is a solution of (3.1). It follows from (3.5) that V is minimal. □

We have implemented this result in the following algorithm to determine a^* for a given $b < 1$. Starting with an estimated value A for a^* and taking $U_0 = 0$, we compute U_n by means of (3.4). The integration is done by dividing the interval $[0, a]$ into N equal subintervals; we use the IMSL library subroutine ICSCCU (cubic spline interpolation in single precision) to interpolate the function U_{n-1} at the $N+1$ subdivision points, the subroutine ICSEVU (cubic spline evaluation in single precision) to evaluate U_{n-1} in the integrand, and the subroutine DCADRE (single precision Romberg integration) to do the integration. In this way we obtain $U_n(x)$ at the points $x = \xi_i$, $i = 2, \cdots, N+1$ of the subdivision. If $U_n(a) < c$ and $\max\{U_n(\xi_i) - U_{n-1}(\xi_i): i = 2, 3, \cdots, N+1\} < 0.5 \times 10^{-d}$ (for a desired accuracy of d decimal places), we claim that u exists globally; if $U_n(a) \geqq c$, we claim that u quenches. In the latter case, we decrease A by a small positive amount, to obtain a new estimate of a^*, and repeat the procedure, until we find that u exists globally. We then use the method of bisection to determine a value a^{**} such that u exists globally for $a \leqq a^{**}$, but u quenches for $a > a^{**}$. We then claim that $a^* = a^{**}$ to the accuracy prescribed.

4. Quenching time. It remains to obtain an upper bound for the quenching time T. Consider the singular Sturm-Liouville problem

$$-x^{-b}(x^b w')' = \lambda^2 w, \qquad x \in (0, a), \tag{4.1a}$$

$$w(0) = 0, \qquad w'(a) = 0. \tag{4.1b}$$

The general solution of (4.1a) is $w(x) = x^\nu[c_1 J_\nu(\lambda x) + c_2 Y_\nu(\lambda x)]$, where c_1 and c_2 are arbitrary constants, $\lambda > 0$, and J_ν and Y_ν are Bessel functions of the first and second kind, respectively, with $\nu = \frac{1}{2}(1-b)$ (cf. [15, § 9.1]). The eigenvalues λ^2 of (4.1a, b) are found from the equation $J_{\nu-1}(\lambda a) = 0$. If $j_{\nu-1}$ denotes the first positive zero of $J_{\nu-1}$, then the smallest positive eigenvalue is $\lambda^2 = (j_{\nu-1}/a)^2$; the corresponding eigenfunction is $x^\nu J_\nu(\lambda x)$.

Any function z that satisfies the differential inequality

$$Lz \geqq -f(z), \qquad (x, t) \in \Omega, \tag{4.2}$$

and the initial and boundary conditions $z(x, 0) = 0$, $z(0, t) = 0$, $z_x(a, t) = 0$, is a lower bound for u, by virtue of the strong maximum principle and the parabolic version of Hopf's lemma. We will seek a lower bound of the form

$$z(x, t) = x^\nu J_\nu(j_{\nu-1} x/a) g(t).$$

According to (4.2), we have

$$g'(t) + \left(\frac{j_{\nu-1}}{a}\right)^2 g(t) \leqq \frac{f(x^\nu J_\nu(j_{\nu-1}x/a)g(t))}{x^\nu J_\nu(j_{\nu-1}x/a)}. \tag{4.3}$$

The expression in the left member is independent of x. Since we are looking for a lower bound, we may take the infimum of the expression in the right member with

respect to x; the infimum exists, because $\nu > 0$. Let $G(g(t))$ satisfy the inequality

$$G(g(t)) \leqq \inf\left\{\frac{f(x^\nu J_\nu(j_{\nu-1}x/a)g(t))}{x^\nu J_\nu(j_{\nu-1}x/a)} : x \in [0, a]\right\}.$$

Then we can determine g as the solution of the initial value problem

$$g'(t) + (j_{\nu-1}/a)^2 g(t) = G(g(t)), \quad t > 0, \quad g(0) = 0. \tag{4.4}$$

Here, the initial condition comes from $u(x, 0) = 0$. As the function $x^\nu J_\nu(j_{\nu-1}x/a)$ is nondecreasing on $[0, a]$, it attains its maximum at $x = a$. Denoting this maximum by m, we have $m = a^\nu J_\nu(j_{\nu-1})$. Thus, an upper bound t_1 for the quenching time is found from the equation

$$mg(t_1) = c. \tag{4.5}$$

5. Example. We illustrate the results of the previous sections for $f(u) = (1-u)^{-1}$.

By Theorem 2, u quenches as $u(a, t)$ tends to 1. In this case, the inequality (4.3) is

$$g'(t) + \left(\frac{j_{\nu-1}}{a}\right)^2 g(t) \leqq \frac{1}{x^\nu J_\nu(j_{\nu-1}x/a)[1 - x^\nu J_\nu(j_{\nu-1}x/a)g(t)]}.$$

It follows from (4.4) and elementary calculus that

$$\frac{g'(t)}{g(t)} + \left(\frac{j_{\nu-1}}{a}\right)^2 = \begin{cases} [mg(t)(1-mg(t))]^{-1}, & 0 < g(t) \leqq (2m)^{-1}, \\ 4, & (2m)^{-1} < g(t) \leqq m^{-1}, \end{cases} \tag{5.1}$$

where $g(0) = 0$. The differential equations can be integrated by a separation of variables. Let t_0 denote the time when $g(t_0) = (2m)^{-1}$. Integrating the first of the equations (5.1) from zero to t_0, we obtain

$$t_0 = \frac{a}{j_{\nu-1}}\left[4 - \left(\frac{j_{\nu-1}}{a}\right)^2\right]^{-1/2} \tan^{-1}\left\{\frac{j_{\nu-1}}{a}\left[4 - \left(\frac{j_{\nu-1}}{a}\right)^2\right]^{-1/2}\right\} - \frac{1}{2}\left(\frac{a}{j_{\nu-1}}\right)^2 \ln\left[1 - \left(\frac{j_{\nu-1}}{2a}\right)^2\right]. \tag{5.2}$$

Next, integrating the second of the equations (5.1) from t_0, we obtain

$$g(t) = (2m)^{-1} \exp\left\{\left[4 - \left(\frac{j_{\nu-1}}{a}\right)^2\right](t - t_0)\right\}. \tag{5.3}$$

According to (4.5), an upper bound t_1 of the quenching time is given by $mg(t_1) = 1$. Using (5.3), we find $\exp\{[4 - (j_{\nu-1}/a)^2](t_1 - t_0)\} = 2$, from which we obtain

$$t_1 = t_0 + \frac{\ln 2}{4 - (j_{\nu-1}/a)^2}. \tag{5.4}$$

We deduce from (5.2) that $4 - (j_{\nu-1}/a)^2 > 0$. Hence, u quenches when $a > \frac{1}{2}j_{\nu-1}$. In particular, when $b = 0$, $\nu = \frac{1}{2}$ and $J_{-1/2}(z) = [2/(\pi z)]^{1/2} \cos z$, so $j_{-1/2} = \frac{1}{2}\pi$. Thus, u

TABLE 1
Critical length a^* for four values of b.

b	ν	$\frac{1}{2}j_{\nu-1}$	a^*
0.40000	0.30000	0.58570	0.57840
0.00000	0.50000	0.78540	0.76515
−0.40000	0.70000	0.96140	0.92314
−1.00000	1.00000	1.20241	1.12927

quenches if $a > \frac{1}{4}\pi$. This result agrees with Walter's conclusion [14] that $l^* \leqq \frac{1}{2}\pi$ for (1.1a, b). The solution quenches for $a = \frac{1}{2}\pi$; (5.2) and (5.4) give the following estimate for the quenching time: $t_1 = \pi/6\sqrt{3} + \ln(2^{4/3}/\sqrt{3}) = 0.67719$ (to five decimal places).

Using the procedure of § 3, we have determined a^* to five decimal places for various values of b. We started the algorithm with the estimate $A = \frac{1}{2}j_{\nu-1} - 0.1$ of a^*, since we already knew that u quenched if $a > \frac{1}{2}j_{\nu-1}$. The results are given in Table 1.

We note that $2a^* = 1.5303$ (to four decimal places) if $b = 0$, in agreement with the result of Acker and Walter [1] for the initial boundary value problem (1.1a, b).

Acknowledgment. The authors thank Professor Man Kam Kwong for helpful discussions.

REFERENCES

[1] A. ACKER AND W. WALTER, *The quenching problem for nonlinear parabolic differential equations*, Lecture Notes in Mathematics 564, Springer-Verlag, New York 1976, pp. 1-12.

[2] ———, *On the global existence of solutions of parabolic differential equations with a singular nonlinear term*, Nonlinear Anal.: Theory, Meth. Appl., 2 (1978), pp. 499-505.

[3] V. ALEXIADES, *A singular parabolic initial-boundary value problem in a noncylindrical domain*, SIAM J. Math. Anal., 11 (1980), pp. 348-357.

[4] ——, *Generalized axially symmetric heat potentials and singular parabolic initial boundary value problems*, Arch. Rational Mech. Anal., 79 (1982), pp. 325-350.

[5] V. ALEXIADES AND C. Y. CHAN, *A singular Fourier problem with nonlinear radiation in a noncylindrical domain*, Nonlinear Anal.: Theory, Meth. Appl., 5 (1981), pp. 835-844.

[6] H. BREZIS, W. ROSENKRANTZ, B. SINGER, AND P. D. LAX, *On a degenerate elliptic-parabolic equation occurring in the theory of probability*, Comm. Pure Appl. Math., 24 (1971), pp. 395-416.

[7] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[8] H. KAWARADA, *On solutions of initial-boundary problem for* $u_t = u_{xx} + (1-u)^{-1}$, Publ. Res. Inst. Math. Sci., 10 (1975), pp. 729-736.

[9] J. LAMPERTI, *A new class of probability theorems*, J. Math. Mech., 11 (1962), pp. 749-772.

[10] H. A. LEVINE, *The phenomenon of quenching: A survey*, in Trends in the Theory and Practice of Non-linear Analysis, V. Lakshmikantham, ed., Elsevier Science, New York, 1985, pp. 275-286.

[11] H. A. LEVINE AND J. T. MONTGOMERY, *The quenching of solutions of some nonlinear parabolic equations*, SIAM J. Math. Anal., 11 (1980), pp. 842-847.

[12] G. N. POLOZHIY, *Equations of Mathematical Physics*, Hayden, New York, 1967.

[13] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[14] W. WALTER, *Parabolic differential equations with a singular nonlinear term*, Funkcial. Ekvac., 19 (1976), pp. 271-277.

[15] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series, Vol. 55, Washington, DC, 1964.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 567-581, May 1989

005

GLOBAL SOLUTIONS AND LONG-TIME BEHAVIOR TO A FITZHUGH–NAGUMO MYELINATED AXON MODEL*

PEI-LI CHEN† AND JONATHAN BELL†‡

Abstract. A class of problems of the following form is considered:

$$u_t = u_{xx} - gu, \qquad x \in \mathbb{R}\backslash Z,$$
$$u_t = u_x(n+, t) - u_x(n-, t) + f(u) - w, \qquad x = n \in Z,$$
$$w_t = \sigma u - \gamma w, \qquad x = n \in Z.$$

In particular, the existence of solutions to the Cauchy problem is proved and the time evolution of solutions to the problem is studied. Such a system models an infinite, myelinated axon with discrete, excitable nodes spaced unit distance apart, and the model dynamics are of Fitzhugh-Nagumo type.

Key words. myelinated axon, Fitzhugh-Nagumo, reaction-diffusion equation

AMS(MOS) subject classifications. 35A05, B40, 92A05

1. Introduction. There has been considerable analytical work done on axon models. Much of this work has been done on models representing an unmyelinated axon that leads to a reaction-diffusion equation, for conservation of current, coupled to one or more ordinary differential equations. A particular form of model that has proved tractable to work with is the so-called Fitzhugh–Nagumo model (see, e.g., [7], [9] and references therein). Particular questions addressed analytically have been the existence and stability of action potentials (traveling wave solutions), as well as various long-time behavior of the solutions.

Myelinated axons, which are much more prevalent in human anatomy than unmyelinated axons, have also been modeled, but analytically such models have been studied much less. Some asymptotic behavior has been considered in the simplest models in [2]–[4]. In such axons their membranes are wrapped in a sheath of lipoprotein, which insulates most of the axon. There are, however, intervals along the axon where gaps, called nodes of Ranvier, appear in the sheathing. This allows a conducting path to the excitable membrane of the axon. Myelination allows the axon to conduct pulses by exciting only a small percentage of membrane, thus permitting transmission at greatly reduced energy expenditure and higher speeds than comparably sized unmyelinated axons.

In this paper we analyze a myelinated axon version of the Fitzhugh–Nagumo model having the following form:

$$u_t = u_{xx} - gu, \qquad x \in \mathbb{R}\backslash Z, \tag{1.1}$$
$$u_t = [u_x]_n + f(u) - w, \qquad x = n \in Z, \tag{1.2}$$
$$w_t = \sigma u - \gamma w, \qquad x = n \in Z. \tag{1.3}$$

Here u represents the transmembrane potential, and w is a "recovery" variable usually associated with long timescale ionic processes. The first equation represents diffusion

* Received by the editors October 26, 1987; accepted for publication (in revised form) July 27, 1988.
† Department of Mathematics, State University of New York, Buffalo, New York 14214.
‡ The research of this author was partially supported by National Science Foundation grant DMS-8615739.

of potential in the myelinated segments, while equations (1.2)-(1.3) represent the dynamics at the nodes. The myelinated segments are scaled to have unit length, and the nodes are located at the integer values along the axis. Here $[u_x]_n$ means $\lim_{\varepsilon\to 0+}\{u_x(n+\varepsilon, t)-u_x(n-\varepsilon, t)\}$, and $f(u)$ is a current-voltage relation to be specified. Parameters g, σ, and γ are positive constants.

In § 2 we prove the existence of solutions to the Cauchy problem for (1.1)-(1.3) via Hilbert space methods. The arguments used in § 2 are similar to those used by Cosner in [5], who has studied a boundary value problem without the recovery process being included. In § 3 we study the long-time behavior of a class of problems of the form (1.1)-(1.3). In particular, we show via Lyapunov techniques when $u\to 0$ as $t\to\infty$ and when $u\to Q(x)\not\equiv 0$ as $t\to\infty$, where $Q(x)$ represents an excited state of the axon.

2. Existence. Consider problem (1.1)-(1.3) with $(u, u(\cdot, t), w)_{t=0}=(\varphi(x), \psi, h)$, where $\varphi(x)$ is continuous and bounded on $\mathbb{R}$ and $\lim_{x\to n}\varphi(x)=\psi(n)$. Here g, σ, and γ are fixed positive constants and $f(u)$ is assumed to be extendible to a function on $\mathbb{C}$ which, if viewed as a function of two variables, is considered C^1. In the usual FitzHugh-Nagumo case, f is taken to be the cubic $f(u)=bu(1-u)(u-a)$, $b>0$, and $a\in(0,1)$. Below our approach is to set up the appropriate machinery and show that the hypotheses of the Sobolev Existence Theorem are satisfied.

We define a linear operator $A: X\to H$ by

$$A(u, v, w)=(-u_{xx}+u, ([-u_x]_j+v^j)_Z, (w^j)_Z), \qquad u(j, \cdot)=v^j.$$

Here

$$H=\{(p, q, r): p\in L^2(\mathbb{R}\backslash Z), q\in l^2(\mathbb{R}), r\in l^2(\mathbb{R})\}$$

with the inner product

$$\langle(p_1, q_1, r_1), (p_2, q_2, r_2)\rangle_H=\int_{\mathbb{R}\backslash Z} p_1\bar{p}_2+\Sigma_Z(q_1^i\bar{q}_2^i+r_1^i\bar{r}_2^i)$$

and

$$X=\{(p, q, r)\in H: p', p''\in L^2(\mathbb{R}\backslash Z), p' \text{ is absolutely continuous on } (j-1, j) \text{ for each } j\in Z, ([p']_j)_Z\in l^2, \lim_{x\to j} p(x)=q^j\}.$$

H is a complete complex Hilbert space and X is dense in H. For the analysis we need another subspace in H, namely

$$Y=\{(p, q, r)\in H: p'\in L^2(\mathbb{R}\backslash Z), p \text{ is absolutely continuous on } (j-1, j) \text{ for each } j\in Z, \lim_{x\to j} p(x)=q^j\}$$

with the inner product

$$\langle(p_1, q_1, r_1), (p_2, q_2, r_2)\rangle_Y=\int_{\mathbb{R}\backslash Z}(p_1\bar{p}_2+p_1'\bar{p}_2')+\Sigma_Z(q_1^j\bar{q}_2^j+r_1^j\bar{r}_2^j).$$

It follows that

$$W\subset Y\subset H$$

with

$$\|(u, v, w)\|_H\leqq\|(u, v, w)\|_Y \quad \text{for any } (u, v, w)\in Y.$$

We first show that A has a bounded inverse, so that A is closed (since dom $A = X$ is dense in H). This is equivalent to solving

$$A(u, v, w) = (p, q, r) \in H, \qquad (u, v, w) \in X \tag{2.1}$$

with $\|(u, v, w)\|_H \leqq K\|(p, q, r)\|_H$ for some constant K.

To solve (2.1) we first obtain a generalized solution in Y and then show that such a solution must belong to X.

When we choose a $(p, q, r) \in H$, for any $(x, y, z) \in Y$, since

$$\langle (x, y, z), (p, q, r)\rangle_H \leqq \|(x, y, z)\|_H \|(p, q, r)\|_H \leqq \|(x, y, z)\|_Y \|(p, q, r)\|_H,$$

$\langle (x, y, z), (p, q, r)\rangle_H$ defines a bounded linear functional on Y, so by the Riesz Representation Theorem, there exists a unique $(u, v, w) \in Y$ such that

$$\langle (x, y, z), (u, v, w)\rangle_Y = \langle (x, y, z), (p, q, r)\rangle_H. \tag{2.2}$$

Thus, for any given $(p, q, r) \in H$, we define $B: H \to Y \subset H$ by

$$B(p, q, r) = (u, v, w).$$

By (2.2),

$$\begin{aligned} \|B(p, q, r)\|_Y^2 &= \langle (u, v, w), (u, v, w)\rangle_Y = \langle (u, v, w), (p, q, r)\rangle_H \\ &\leqq \|(u, v, w)\|_Y \|(p, q, r)\|_H \\ &= \|B(p, q, r)\|_Y \|(p, q, r)\|_H. \end{aligned}$$

Therefore

$$\|B(p, q, r)\|_Y \leqq \|(p, q, r)\|_H,$$

$\|B\| \leqq 1$; that is, B is bounded.

It remains to show that B maps H into X and that any solution (u, v, w) to (2.2) in X satisfies $A(u, v, w) = (p, q, r) \in H$.

To prove B maps H into X, the analysis in [6, § I.15] can be used to conclude that $u|_{(a,b)} \in H^2[(a, b)]$ for any a, b with $j-1 < a < b < j$, $j \in Z$. This occurs provided $(u, v, w) \in Y$ satisfies (2.2) for a given $(p, q, r) \in H$, and there exists a constant C such that for any $\varphi \in C_0^\infty[(j-1, j)]$, we have

$$\left| \int_{(j-1,j)} \varphi' \bar{u}' \right| \leqq C\|\varphi\|_{L^2[(j-1,j)]}. \tag{2.3}$$

The inequality (2.3) is essentially the inequality (15.6) in [6, § I.15].

Choose $\varphi \in C_0^\infty(\mathbb{R})$ with supp $\varphi \subset [a, b] \subset (j-1, j)$, and let $\psi = (\psi^k)_Z$, $\eta = (\eta^k)_Z$ be such that $\psi^k = \eta^k = 0$ for all k; then

$$(\varphi, \psi, \eta) \in Y,$$

and (2.2) yields

$$\int_{(j-1,j)} \varphi' \bar{u}' = \int_{(j-1,j)} \varphi(\bar{p} - \bar{u}). \tag{2.4}$$

Hence, by Hölder's inequality,

$$\left| \int_{(j-1,j)} \varphi' \bar{u}' \right| \leqq (\|(p, q, r)\|_H + \|B(p, q, r)\|_H)\|\varphi\|_{L^2[(j-1,j)]}$$

since (p, q, r) is given and $B: H \to H$ is bounded, this establishes (2.3). Therefore, $u''|_{(a,b)} \in L^2(a, b)$ for any $[a, b] \subset (j-1, j)$. By taking complex conjugates in (2.4), we have

$$-\int_{(a,b)} u''\bar{\varphi} = \overline{\int_{(a,b)} \varphi'\bar{u}'} = \int_{(a,b)} \bar{\varphi}(p-u)$$

for any $\varphi \in C_0^\infty[(a, b)]$. By choosing a suitable φ and by (2.2) we also can conclude that $u'', u' \in L^2(\mathbb{R})$.

Since C_0^∞ is dense in L^2, for any $h \in L^2[(a, b)]$,

$$\int_{(a,b)} (-u'')\bar{h} = \int_{(a,b)} (p-u)\bar{h};$$

therefore,

$$-u'' = p-u \quad \text{or} \quad -u''+u = p \quad \text{on } (a, b);$$

hence

$$-u''+u = p \quad \text{on } (j-1, j)$$
$$-u''+u = p \quad \text{on } \mathbb{R}\backslash Z;$$

this is the first equation of (2.1).

Since $L^2[(j-1, j)] \subset L^1[(j-1, j)]$ it follows that u' is absolutely continuous on $(j-1, j)$ and that

$$u'(b)-u'(a) = \int_{(a,b)} u'' \quad \text{for any } (a, b) \subset (j-1, j).$$

Let $a \to j-1$ or $b \to j$ to see that $u'(j-1+)$ and $u'(j-)$ are well defined. Next we verify the last two equations in (2.1).

Let $(w_j, y_j, 0)$ be such that $y_j^k = \delta_{jk}$, and $w_j' = (1/h)\chi_{(j-h,j)}$ on $(j-h, j)$, $w_j' = -(1/h)\chi_{(j,j+h)}$ on $(j, j+h)$, and $w' = 0$ otherwise, where $h > 0$ is arbitrary. Then $w_j = (1/h)(x-j+h)$ on $(j-h, j)$, $w_j = -(1/h)(x-j-h)$ on $(j, j+h)$, and $w_j = 0$ otherwise. It follows that $(w_j, y_j, 0) \in Y$, and (2.2) yields

$$\left(\frac{1}{h}\right)\int_{(j-h,j)} u' + \left(\frac{1}{h}\right)\int_{(j-h,j)} (x-j+h)u - \left(\frac{1}{h}\right)\int_{(j,j+h)} u' - \left(\frac{1}{h}\right)\int_{(j,j+h)} (x-j-h)u + v^j$$

(2.5)

$$= \left(\frac{1}{h}\right)\int_{(j-h,j)} (x-j+h)p - \left(\frac{1}{h}\right)\int_{(j,j+h)} (x-j-h)p + q^j.$$

Noting that

$$\left(\frac{1}{h}\right)\int_{(j-h,j)} u' = \left(\frac{1}{h}\right)[u(j)-u(j-h)] \to u'(j-),$$

$$\left(\frac{1}{h}\right)\int_{(j,j+h)} u' \to u'(j+) \quad \text{as } h \to 0,$$

and $|(1/h)\int_{(j-h,j)} (x-j+h)w| \leqq \int_{(j-h,j)} |w| \to 0$ as $h \to 0$ for any $w \in l^2(\mathbb{R})$, we may let $h \to 0$ in (2.5), to obtain the equation

$$u'(j-)-u'(j+)+v^j = q^j,$$

which is the second equation of (2.1).

Choose $(0, 0, z_j) \in Y$, with $z_j^k = \delta_{jk}$, substitute this into (2.2), i.e.,

$$\langle (0, 0, z_j), (u, v, w) \rangle_Y = \langle (0, 0, z_j), (p, q, r) \rangle_H ;$$

hence $w^j = r^j$, and so we have the third equation.

Thus B is a map from H to X, and for any $(p, q, r) \in H$, $B(p, q, r) = (u, v, w)$, (u, v, w) satisfies (2.3) and $AB(p, q, r) = (p, q, r)$ and also by the construction we have $BA(u, v, w) = (u, v, w)$ for any $(u, v, w) \subset X$.

Since $B = A^{-1}$ is a bounded operator defined on all of H, $X = \operatorname{dom} A$ is dense in H, B is linear, and $\|B\|_H \neq 0$, so $\|B\|_X = \|B\|_H$ and $\|A\|_X = \|B\|_X^{-1} = \|B\|_H^{-1}$. A is a closed, bounded operator, so we have proved the following lemma.

LEMMA 1. *The operator A defined by $A(u, v, w) = (-u_{xx} + u, (-[u_x]_j + v^j)_Z, (w^j)_Z)$ with $\operatorname{dom} A = X$ is a closed operator from X to H with $A^{-1} : H \to H$ bounded.*

To apply semigroup theory to our nonlinear problem, we need to show that the operator A satisfies

$$\|(\lambda - A)^{-1}\| \leqq C/(1 + |\lambda|) \quad \text{for } \operatorname{Re} \lambda \leqq 0 \tag{2.6}$$

for some constant $C > 0$, where $\| \cdot \|$ denotes the operator norm on $B(H)$. Let $\theta(A) \subset \mathbb{C}$ be given by

$$\theta(A) = \{ \langle A(u, v, w), (u, v, w) \rangle_H : (u, v, w) \in \operatorname{dom} A, \|(u, v, w)\|_H = 1 \}$$

and let

$$\Gamma = \operatorname{cl}(\theta(A)), \qquad \Delta = \mathbb{C} \backslash \Gamma.$$

Choosing $\|(u, v, w)\|_H = 1$, we have

$$\begin{aligned} \langle A(u, v, w), (u, v, w) \rangle_H &= \int_{\mathbb{R} \backslash Z} (u'' + u)\bar{u} + \Sigma_Z ((-[u']_j + v^j)\bar{v}^j + w^j \bar{w}^j) \\ &= \int_{\mathbb{R} \backslash Z} (|u'|^2 + |u|^2) + \Sigma_Z u'(j+)\bar{u}(j+) - \Sigma_Z u'(j-)\bar{u}(j-) \\ &\quad - \Sigma_Z [u']_j \bar{v}^j + \Sigma_Z (|v^j|^2 + |w^j|^2); \end{aligned}$$

here $u' = u_x$, $u'' = u_{xx}$. Since

$$u(j+) = u(j-) = v^j \quad \text{for } (u, v, w) \in X$$

and

$$u'(j+) - u'(j-) = [u']_j,$$

then

$$\begin{aligned} \langle A(u, v, w), (u, v, w) \rangle_H &= \int_{\mathbb{R} \backslash Z} (|u'|^2 + |u|^2) + \Sigma_Z (|v^j|^2 + |w^j|^2) \\ &= \|(u, v, w)\|_Y^2 \geqq \|(u, v, w)\|_H^2 = 1. \end{aligned}$$

Now $\theta(A) \subset [1, \infty)$ and $\Gamma \subset [1, \infty)$, and $\Delta = \mathbb{C}/\Gamma$ is connected in $\mathbb{C}$ since $A^{-1} \in B(H)$ and A has deficiency zero on Δ. Later we will use the following lemma, which appears in [5], and which is a special case of Theorem 3.2 of [8, Chap. V].

LEMMA 2. *Suppose that A is closed and Δ is connected. For $\lambda \in \Delta$, $A-\lambda$ has nullity zero and constant deficiency. If the deficiency of $A-\lambda$ is zero, then Δ is contained in the resolvent set of A and*

$$\|(\lambda - A)^{-1}\| \leqq 1/\operatorname{dist}(\lambda, \Gamma) \tag{2.7}$$

where $\operatorname{dist}(\lambda, \Gamma)$ *is the distance from* λ *to* Γ *in* $\mathbb{C}$.

By (2.7) for $\operatorname{Re}\lambda \leqq 0$, if we suppose $\lambda = a + b\vec{i}$, then we can easily conclude that $\operatorname{dist}(\lambda, \Gamma) \geqq (1+\lambda)/2$; hence A satisfies (2.6).

To prove the solution of (1.1)-(1.3) exists, we rewrite (1.1)-(1.3) as follows:

$$\begin{aligned} &u_t - u_{xx} + u = (1-g)u, \\ &v_t^j - [u_x]_j + v^j = v^j + f(v^j) - w^j, \\ &w_t^j + w^j = (1-\gamma)w^j + \sigma v^j, \\ &(u, v, w)|_{t=0} = (\varphi, \psi, \eta), \end{aligned} \tag{2.8}$$

and let

$$F(t, (u, v, w)) = ((1-g)u, (v^j + f(v^j) - w^j)_Z, ((1-\gamma)w^j + \sigma v^j)_Z).$$

Then (2.8) can be written abstractly as

$$\frac{d}{dt}(u, v, w) + A(u, v, w) = F(t, (u, v, w)), \tag{2.9}$$

$$(u, v, w)|_{t=0} = (\varphi, \psi, \eta). \tag{2.10}$$

Here we require that

$$(\varphi, \psi, \eta) \in \operatorname{dom} A = X. \tag{2.11}$$

Next, we use the following lemma (see the Sobolev theorem in [6]) to prove (2.9)-(2.11) has a solution.

LEMMA 3. *Let A be a closed linear operator on a Banach space E such that* (2.7) *holds. Suppose that $F(t, p)$ is a function on $[0, T_0] \times E$ such that for some constants $\alpha, \eta \in (0, 1)$ and for any $R > 0$ there exists a constant $C(R)$ for which*

$$\|F(t_1, A^{-\alpha}p_1) - F(t_2, A^{-\alpha}p_2)\|_E \leqq C(R)[|t_1 - t_2|^\eta + \|p_1 - p_2\|_E] \tag{2.12}$$

for all $t_1, t_2 \in [0, T_0]$, $p_1, p_2 \in E$ with $\|p_1\|_E, \|p\|_E < R$. Then for any $p_0 \in \operatorname{dom}(A)$ and each $R > \|A^\alpha p_0\|_E$, there exists a $t^* = t^*(R, \|A^\alpha p_0\|_E) > 0$ such that the problem

$$\frac{dp}{dt} + Ap = F(t, p), \qquad p(0) = p_0 \tag{2.13}$$

has a unique solution in $[0, t^*]$. Furthermore, if there exists a constant $R' > 0$ such that for any solution p of (2.13) *in $[0, T_1]$, $T_1 \leqq T_0$, we have*

$$\|Ap\|_E < R',$$

then we may choose $R > R'$ and thus apply the local existence assertion to $[0, t^*]$, $[t^*, 2t^*]$, and so on until $[0, T_0]$ is exhausted.

Since the operator A in (2.13) was shown to be closed on X and to satisfy (2.7), we need only establish (2.12) for the function F in (2.9) to conclude the local existence of a solution to (2.1). Let

$$(x_k, y_k, z_k) = A^{-\alpha}(u_k, v_k, w_k) \quad \text{for } k = 1, 2.$$

If $\|(u_k, v_k, w_k)\|_H < R$ for $k = 1, 2$, then

$$|y_k^j| \leqq \|y_k\|_{l^2} \leqq \|(x_k, y_k, z_k)\|_H \leqq \|A^{-\alpha}\| R;$$

so

$$\begin{aligned}
\|(f(y_1^j) - f(y_2^j))_Z\|_{l^2} &\leqq \sup_{|y| \leq \|A^{-\alpha}\| R} |f'(y)| \, \|y_1 - y_2\|_{l^2} \\
&\leqq C_0(R) \|(x_1 - x_2, y_1 - y_2, z_1 - z_2)\|_H \\
&\leqq C_0(R) \|A^{-\alpha}\| \, \|(u_1, v_1, w_1) - (u_2, v_2, w_2)\|_H,
\end{aligned}$$

$$\begin{aligned}
&\|(y_1^j - y_2^j + f(y_1^j) - f(y_2^j) - (z_1^j - z_2^j))_Z\|_{l^2} \\
&\qquad \leqq (C_0(R) + 2) \|A^{-\alpha}\| \, \|(u_1, v_1, w_1) - (u_2, v_2, w_2)\|_H \qquad (2.14) \\
&\qquad = C_1(R) \|(u_1, v_1, w_1) - (u_2, v_2, w_2)\|_H.
\end{aligned}$$

Then

$$\begin{aligned}
&\|F(t_1, A^{-\alpha}(u_1, v_1, w_1)) - F(t_2, A^{-\alpha}(u_2, v_2, w_2))\|_H \\
&\quad = \|F(t_1, (x_1, y_1, z_1)) - F(t_2, (x_2, y_2, z_2))\|_H \\
&\quad = \|(1 - g)(x_1 - x_2), (y_1^j - y_2^j + f(y_1^j) - f(y_2^j) - (z_1^j - z_2^j))_Z, \\
&\qquad\qquad ((1 - \gamma)(z_1^j - z_2^j) + \sigma(y_1^j - y_2^j))_Z\|_H \qquad (2.15) \\
&\quad \leqq |1 - g| \, \|(x_1 - x_2, 0, 0)\|_H \\
&\qquad + \|(0, (y_1^j - y_2^j + f(y_1^j) - f(y_2^j) - (z_1^j - z_2^j))_Z, 0)\|_H \\
&\qquad + |1 - \gamma| \, \|(0, 0, z_1^j - z_2^j)\|_H + \sigma \|(0, 0, y_1^j - y_2^j)_Z\|_H,
\end{aligned}$$

since

$$\begin{aligned}
\|(x, 0, 0)\|_H &\leqq \|(x, y, z)\|_H, \\
\|(0, y, 0)\|_H &\leqq \|(x, y, z)\|_H, \\
\|(0, 0, z)\|_H &\leqq \|(x, y, z)\|_H.
\end{aligned}$$

From (2.14) and (2.15)

$$\begin{aligned}
&\|F(t_1, A^{-\alpha}(u_1, v_1, w_1)) - F(t_2, A^{-\alpha}(u_2, v_2, w_2))\|_H \\
&\quad \leqq (|1 - g| + C_1(R) + |1 + \gamma| + \sigma) \|A^{-\alpha}\| \, \|(u_1, v_1, w_2) - (u_2, v_2, w_2)\|_H. \qquad (2.16)
\end{aligned}$$

Hence we have the local existence of a unique solution to (2.9)-(2.11).

To obtain the global existence we must bound $\|A(u, v, w)\|_H$ for any solution $P = (u, v, w)$ to (2.9)-(2.11). It is sufficient to bound $\|(u, v, w)\|_H$, and $\|(u_t, v_t, w_t)\|$, since, if $\|(u, v, w)\|_H \leqq M$, then $|v^j| \leqq M$ and

$$|f(v^j)| \leqq \sup_{|y| \leqq M} |f'(y)| |v^j|.$$

Hence $f(v^j)$ is bounded provided $\sup_{|y| \leqq M} |f'(y)|$ is bounded.

Now the idea is to construct a Lyapunov function and prove that $\|(u, v, w)\|_H$ and $\|(u_t, v_t, w_t)\|$ are bounded. Let

$$E(t) = (\|(u, v, w)\|_H^2 + \|(u, v, w)_t\|_H^2)/2.$$

If we only consider that u, $(v^j)_Z$, $(w^j)_Z$ are real functions, then

$$
\begin{aligned}
E'(t) &= \int_{\mathbb{R}} (uu_t + u_t u_{tt}) + \Sigma_Z (v_t^j v_{tt}^j + v_t^j v^j + w^j w_t^j + w_t^j w_{tt}^j) \\
&= \int_{\mathbb{R}} u(u_{xx} - gu) + u_t(u_{xxt} - gu_t) \\
&\quad + \Sigma_Z \{ v_t^j [u_{xt}]_j + f'(v^j)|v_t^j|^2 - w_t^j v_t^j + v^j [u_x]_j + f(v^j) v^j - w^j v^j \\
&\qquad - \gamma |w^j|^2 + \sigma v^j w^j - \gamma |w_t^j|^2 + \sigma v_t^j w_t^j \} \qquad (2.17) \\
&= -\int_{\mathbb{R}} (u_x^2 + gu^2 + u_{tx}^2 + gu_t^2) \\
&\quad + \Sigma_Z \{ f'(v^j)|v_t^j|^2 + f'(\theta v^j)|v^j|^2 + (\sigma - 1) w_t^j v_t^j + (\sigma - 1) v^j w^j \\
&\qquad - \gamma((w^j)^2 + (w_t^j)^2) \} \\
&= I_1 + I_2,
\end{aligned}
$$

where

$$
I_1 = -\int_{\mathbb{R}} (u_x^2 + gu^2 + u_{xt}^2 + gu_t^2),
$$

$$
I_2 = \Sigma_Z \{ f'(v^j)|v_t^j|^2 + (\sigma - 1) w_t^j v_t^j - \gamma |w_t^j|^2 + f'(\theta v^j)|v^j|^2 + (\sigma - 1) v^j w^j - \gamma |w^j|^2 \}.
$$

Suppose there exists a positive constant $k_1 > 0$ such that

$$
\sup_{y \in \mathbb{R}} (f'(y)) = k_1; \qquad (2.18)
$$

then

$$
I_2 \leqq \Sigma_Z \{ k_1 (|v^j|^2 + |v_t^j|^2) + (\sigma - 1)(v^j w^j + v_t^j w_t^j) \}.
$$

By Cauchy's inequality, we can find another constant $k_2 > 0$, which makes

$$
I_2 \leqq k_2 \Sigma_Z \{ |w_t^j|^2 + |v_t^j|^2 + |w^j|^2 + |v^j|^2 \};
$$

hence

$$
E'(t) \leqq k_2 E(t),
$$

which implies

$$
E(t) \leqq A E^{k_2 t} \qquad (2.19)
$$

for some constant A.

Since we have already proved the local existence to (2.9)-(2.11) in $[0, T_1]$ for $T_1 \leqq T_0$, by (2.19) we know that $\|(u, v, w)\|_H$, $\|(u, v, w)_t\|_H$ are bounded in $[0, T_0]$. Hence there exists an $R' > 0$ such that $\|A(u, v, w)\|_H < R'$ for any solution of (2.9)-(2.11) in $[0, T_0]$. By using Lemma 3, (2.9)-(2.10) have a unique global solution in $[0, T_0]$. Since T_0 is arbitrary, we have the following theorem.

THEOREM. *If* $(\varphi_0, \psi_0, \eta_0) \in X$ *and condition* (2.18) *holds, then the problem* (2.1) *has a unique solution* $(u, v, w) \in X$ *for all* $t > 0$.

3. Long-time behavior of solutions. Suppose $u(x, t)$, $w^1(n, t)$ is the solution of the following system:

$$
u_t = u_{xx} - gu, \qquad x \in \mathbb{R} \setminus Z, \qquad (3.1)
$$

$$
u_t = [u_x]_n + F(u) - w^1, \qquad x = n \in Z, \qquad (3.2)
$$

$$
w_t^1 = \sigma u - \gamma w^1, \qquad x = n \in Z. \qquad (3.3)
$$

Here g, σ, and γ are positive constants and $F \in C^1(\mathbb{R})$ satisfies the condition that there exists an interval $(\alpha, \beta) \in \mathbb{R}$ and a constant $B > 0$, such that $B \leqq F'(y) < 0$ whenever $y \in (\alpha, \beta)$.

Let $Q(x)$, $w^2(n)$, for $x \in \mathbb{R}$ and $n \in Z$, be such that $(Q, (Q^j)_Z, (w^{2j})_Z)$ $(Q^j = \lim_{x \to j} Q(x))$ is a solution to the stationary problem

$$Q_{xx}(x) - gQ(x) = 0, \qquad x \in \mathbb{R} \setminus Z, \tag{3.4}$$

$$[Q_x(n)]_n + F(Q(n)) - w^2 = 0, \qquad x = n \in Z, \tag{3.5}$$

$$\sigma Q(n) - \gamma w^2(n) = 0, \qquad x = n \in Z. \tag{3.6}$$

We discuss the existence of such solutions in the next section.

Define

$$V(x, t) \equiv u(x, t) - Q(x), \qquad x \in \mathbb{R},$$

$$W(n, t) \equiv w^1(n, t) - w^2(n), \qquad n \in Z,$$

and

$$E_0(t) = \int_{\mathbb{R} \setminus Z} (V^2 + V_x^2)\, dx/2,$$

$$E(t) = \left[\int_{\mathbb{R} \setminus Z} (V^2 + aV_x^2)\, dx + \Sigma_Z b(W^2 + \sigma V^2) \right] \Big/ 2,$$

where

$$a = \min \left\{ 1, \frac{\gamma}{\sigma}, \frac{2}{B} \right\} \quad \text{and} \quad b = \frac{ag + 1}{\sigma}.$$

Suppose u, w^1, Q satisfy the following:

(i) $u(x, t)$ is C^3 in $x \in \mathbb{R} \setminus Z$, C^0 in $x \in \mathbb{R}$, and C^1 in $t \in \mathbb{R}^+$; u_t is C^0 in $x \in \mathbb{R}$; $Q(x)$ is C^3 in $x \in \mathbb{R} \setminus Z$, C^0 in $x \in \mathbb{R}$; $w^1(n, t)$ is C^1 in $t \in \mathbb{R}^+$ for each $n \in Z$.

(ii) There exists a $\delta > 0$ such that $-B \leqq F'(y) \leqq -a_1 < 0$ whenever $y \in (\alpha + \delta/2, \beta - \delta/2)$, the steady-state solution $Q(x) \in (\alpha + \delta, \beta - \delta)$ and $u(x, 0)$ satisfies $|u(x, 0) - Q(x)| < \delta/2$.

(iii) $E(t)$, $(d/dt)E(t)$ are uniformly convergent on $\mathbb{R}$, and suppose

$$\{E(0)\}^{1/2} < \delta\sqrt{a}/(2K),$$

where K is a constant satisfying

$$\sup_{x \in \mathbb{R}} |u(x, t) - Q(x)| \leqq K\{E_0(t)\}^{1/2}$$

(see [1, Lemma 5.15]).

In § 2 we have shown that if the initial data lies in X, so does the solution. In this section we desire a bit more smoothness to the solution, as given by (i). We suppose that given smooth enough initial data, that the resulting solution will be sufficiently smooth, then we can conclude the following theorem.

THEOREM. *If $u(x, t)$, $w^1(n, t)$, and $Q(x)$ satisfy the above conditions, then $u(x, t) \to Q(x)$ uniformly on $\mathbb{R}$ as $t \to \infty$, and $w^1(n, t) \to w^2(n)$ uniformly on Z as $t \to \infty$.*

Proof. By (2.1)-(2.3) and (2.4)-(2.5), $V(x, t)$ and $W(n, t)$ satisfy the following equations:

$$V_t = V_{xx} - gV, \qquad x \in \mathbb{R}\backslash Z, \tag{3.7}$$

$$V_t = [V_x]_n + F'(Q+\theta(u-Q))V - W, \qquad x = n \in Z, \tag{3.8}$$

$$W_t = \sigma V - \gamma W, \qquad x = n \in Z \tag{3.9}$$

where $\theta = \theta(x, t) \in (0, 1)$. Now consider $(d/dt)E(t)$:

$$\begin{aligned} E'(t) &= \int_{\mathbb{R}/Z} (VV_t + aV_xV_{xt})\,dx + \Sigma_Z[bWW_t + \sigma bVV_t] \\ &= \Sigma_Z \left\{ \int_n^{n+1} [V(V_{xx} - gV) + aV_x(V_{xx} - gV)_x]\,dx \right. \\ &\qquad \left. + bW(\sigma V - \gamma W) + \sigma bV([V_x]_n + F'(\xi(x))V - W) \right\}, \end{aligned}$$

where (3.1)-(3.3) has been used, and

$$\xi(x) = Q + \theta(u - Q).$$

We can put this in the following form:

$$E'(t) = -\int_{\mathbb{R}\backslash Z} (V_x^2 + \tilde{g}V^2 + aV_{xx}^2 + a\tilde{g}V_x^2)\,dx - \Sigma_Z\{\tilde{g}([V_x]_n, V) + h([V_x]_n, W)\},$$

where

$$\tilde{g}(x, y) = \frac{ax^2}{2} + aF'(\xi)xy + \sigma b(-F'(\xi))y^2,$$

$$h(x, y) = \frac{ax^2}{2} - axy - b\gamma y^2.$$

If $|u(x, t) - Q(x)| \leqq \delta/2$ and $Q(x) \in (a+\delta, b-\delta)$, then by (ii),

$$-B \leqq F'(\xi) \leqq -a_1 < 0;$$

so

$$\begin{aligned} \left(\frac{a}{2}\right)\sigma b(-F'(\xi)) &= \left(\frac{a}{2}\right)(ag+1)(-F'(\xi)) \\ &> \left(\frac{a}{2}\right)(-F'(\xi)). \end{aligned}$$

Since $a = \min\{\gamma/\sigma, 1, 2/B\}$,

$$1 \geqq \frac{aB}{2} \geqq \frac{a(-F'(\xi))}{2};$$

hence

$$\left(\frac{a}{2}\right)\sigma b(-F'(\xi)) \geqq \left[\frac{a(-F'(\xi))}{2}\right]^2;$$

$\tilde{g}(x, y)$ is positive definite, and there exists a $p > 0$ such that

$$\tilde{g}(x, y) \geqq p(x^2 + y^2) \quad \text{for all } (x, y) \in \mathbb{R}.$$

Similarly, since

$$\frac{ab\gamma}{2}=\frac{a\gamma(ag+1)}{2\sigma}$$

$$\geqq\frac{a^2(ag+1)}{2}>\frac{a^2}{4},$$

there exists a $q>0$, such that

$$h(x, y)\geqq q(x^2+y^2).$$

Therefore

$$E'(t)\leqq-\int_{\mathbb{R}\setminus Z}[(ag+1)V_x^2+gV^2+aV_{xx}^2]\,dx-\Sigma_Z\{p(V^2+aV_x^2)+q([V_x]_n^2+W^2)\}$$

$$\leqq-g\int_{\mathbb{R}/Z}(V^2+aV_x^2)\,dx-\Sigma_Z\{pV^2+qW^2\}.$$

Choose a constant $r_1>0$ such that

$$\frac{(ag+1)r_1}{2\sigma}\leqq p,\quad\frac{(ag+1)r_1}{2}\leqq q,\quad\frac{r_1}{2}\leqq g.$$

Then we have

$$E'(t)\leqq-r_1E(t).$$

Now we claim that

$$R(x, t)=|V(x, t)|=|u(x, t)-Q(x)|\leqq\delta/2\quad\text{for all } t\in\mathbb{R}^+.$$

By the initial condition, $R(0, t)<\delta/2$; so suppose there exists a $t^*>0$ such that $R(x_0, t^*)>\delta/2$ for some $x_0\in\mathbb{R}$. Define $T_1=\inf\{t>0\colon R(x, t)>\delta/2 \text{ for some } x\in\mathbb{R}\}$. Since $R(0, t)<\delta/2$ and $R(x_0, t^*)>\delta/2$, this T_1 exists, and $R(x, t)\leqq\delta/2$ whenever $t<T_1$. This means

$$Q(x)+\theta(u(x, t)-Q(x))\in(a+\delta/2, b-\delta/2);$$

so

$$F'(Q(x)+\theta(u(x, t)-Q(x)))\leqq-a_1<0\quad\text{for all } t<T_1.$$

This implies

$$E'(t)\leqq-r_1E(t)\quad\text{for all } t<T_1,$$

and so we have

$$E(T_1)\leqq E(0).$$

On the other hand,

$$\sup_{x\in\mathbb{R}}R(x, T_1)\leqq K\{E_0(T_1)\}^{1/2}$$

$$\leqq K/\sqrt{a}\{E(T_1)\}^{1/2}$$

$$\leqq K/\sqrt{a}\{E(0)\}^{1/2}<\delta/2.$$

This contradicts the assumption $R(x_0, T_1) \geqq \delta/2$. Therefore, such a point (x_0, T_1) does not exist, and so

$$R(x, t) = |u(x, t) - Q(x)| \leqq \delta/2 \quad \text{for all } t \in \mathbb{R}^+, \quad x \in \mathbb{R}.$$

Also

$$F'(Q(x) + \theta(u(x, t) - Q(x))) \leqq -a_1 < 0 \quad \text{for all } x \in \mathbb{R}, \quad t \in \mathbb{R}^+.$$

With

$$\frac{d}{dt} E(t) \leqq -r_1 E(t) \quad \text{for all } t \in \mathbb{R}^+,$$

we have $E(t) \to 0$ as $t \to \infty$, and since $\sup_{x \in \mathbb{R}} |u(x, t) - Q(x)| \leqq K\{E_0(t)\}^{1/2}$, we have $u(x, t) \to Q(x)$ uniformly on $\mathbb{R}$, as $t \to \infty$. Since

$$\frac{ag+1}{2\sigma} |w_1(n, t) - w_2(n)|^2 \leqq \{E(t)\}^{1/2} \quad \text{for all } n \in Z,$$

$$w_1(n, t) \to w_2(n) = \frac{\sigma}{\gamma} Q(n) \quad \text{uniformly on } Z \text{ as } t \to \infty.$$

This completes the proof.

As an example, consider the function

$$f(x) = x(x - \alpha)(1 - x) \quad \text{where } 0 < \alpha < 1 \text{ and } x \in [0, 1].$$

Now

$$f'(x) = -3x^2 + 2(1+\alpha)x - \alpha, \qquad f''(x) = -6x + 2(1+\alpha).$$

The root of $f''(x) = 0$ is $x = (1+\alpha)/3$ and so we have

$$f'\left(\frac{1+\alpha}{3}\right) = \max_{0 \leqq x \leqq 1} f'(x)$$

$$= \frac{1 + \alpha^2 - \alpha}{3} > 0.$$

Also, $f'(0) = -\alpha < 0$, $f'(1) = -1 + \alpha < 0$. The two roots of $f'(x) = 0$ are

$$x_1 = \frac{\alpha}{1 + \alpha + (1 + \alpha^2 - \alpha)^{1/2}}, \qquad x_2 = \frac{1 + \alpha + (1 + \alpha^2 - \alpha)^{1/2}}{3}.$$

If we choose $F(x) = f(x)$ in the theorem, then $F'(x) < 0$ whenever $x \in (0, x_1)$ or $x \in (x_2, 1)$, and, in fact,

$$-\max\{\alpha, 1 - \alpha\} \leqq F'(x) < 0 \quad \text{on } (0, x_1) \cup (x_2, 1).$$

Hence we have the following corollary.

COROLLARY. *If we choose the function* $F(y) = y(y - \alpha)(1 - y)$, $0 < \alpha < 1$ *and* $y \in [0, 1]$, *and for the interval* $(a, b) = (0, x_1)$ *or* $(a, b) = (x_2, 1)$, *then we have* $u(x, t) \to Q(x)$ *uniformly on* $\mathbb{R}$ *as* $t \to \infty$, *and* $w_1(n, t) \to (\sigma/\gamma) Q(n)$ *uniformly on* Z *as* $t \to \infty$.

If $(a, b) = (0, x_1)$, *it is the subthreshold long-time behavior. If* $(a, b) = (x_2, 1)$, *it is the superthreshold long-time behavior.*

4. Steady-state solutions. It remains to show the steady-state solution exists in the intervals $(0, x_1)$ or $(x_2, 1)$.

Considering u as a function of x, then (1.1)-(1.3) becomes

$$u_{xx} - gu = 0, \qquad x \in \mathbb{R} \setminus Z, \tag{4.1}$$

$$[u_x]_n + f(u) - \frac{\sigma}{\gamma} u = 0, \qquad x = n \in Z \tag{4.2}$$

where now we take the case $f(u) = u(u-\alpha)(1-u)$. In the steady-state case w is given by $w(n) = \sigma u(n)/\gamma$.

From (4.1), the general solution is

$$u(x) = [v_{j+1} \sinh \sqrt{g}(x-j) + v_j \sinh \sqrt{g}(j+1-x)]/\sinh\sqrt{g},$$
$$j \leqq x \leqq j+1, \quad j \in Z,$$

where $v_j = u(j)$, $j \in Z$. Letting $G = \sqrt{g}/\sinh(\sqrt{g})$, we have

$$[u_x]_n = G(v_{n+1} - 2\cosh\sqrt{g}\, v_n + v_{n-1}).$$

Substituting this expression in (4.2), we obtain

$$v_{j+1} - 2v_j + v_{j-1} + F(v_j) = 0 \tag{4.3}$$

where $F(v_j) = \{2(1-\cosh\sqrt{g}) + ((v_j - \alpha)(1-v_j) - \sigma/\gamma)/G\}v_j$.

First, we consider $\{v_j\}$ as a constant solution of (4.3), i.e., $v_j \equiv c > 0$ for $j \in Z$.

Then (4.3) becomes

$$\frac{\sigma}{\gamma} + 2G(\cosh\sqrt{g} - 1) = (c-\alpha)(1-c). \tag{4.4}$$

Let $c = x$ in (4.4) and consider the function

$$y(x) = x^2 - (1+\alpha)x + K$$

where $K = \alpha + 2(\cosh\sqrt{g} - 1)G + \sigma/\gamma$. Then the two roots of $y(x) = 0$ are

$$x_{\sigma/\gamma, g}(\alpha) = (1 + \alpha \pm ((1+\alpha)^2 - 4K)^{1/2})/2. \tag{4.5}$$

To have two distinct, real roots, we must have

$$(1+\alpha)^2 > 4K$$

or, equivalently,

$$(1-\alpha)^2 > 4\frac{\sigma}{\gamma} + 8G(\cosh\sqrt{g} - 1). \tag{4.6}$$

If α, σ, and γ satisfy $(1-\alpha)^2 > 4(\sigma/\gamma)$, since $\lim_{x\to 0}(x/\sinh x) = 1$, $\lim_{x\to 0}\cosh x = 1$, by (4.6) we can find a small $g_b > 0$, such that $y(x) = 0$ has two roots C_{1g}, C_{2g} whenever $0 < g < g_b$.

In this case, (4.1)-(4.2) has two solutions:

$$Q_{1g}(x) = C_{1g}R(x), \tag{4.7}$$

$$Q_{2g}(x) = C_{2g}R(x), \tag{4.8}$$

where

$$R(x) = [\sinh\sqrt{g}(x-j) + \sinh\sqrt{g}(j+1-x)]/\sinh\sqrt{g},$$
$$j \leqq x \leqq j+1, \qquad j \in Z$$

and

$$0 < C_{1g} < C_{2g} < 1.$$

Next, we use this result to see whether there is a solution $Q(x)$ of (4.1)-(4.2) satisfying the corollary in § 3.

Example. If we choose $\alpha = 2/3$, $\sigma/\gamma = 1/40$, and g is small enough, there exists a $Q_{2g}(x)$ that lies entirely in $(x_2, 1)$.

To show this, we see that

$$\begin{aligned} x_2|_{\alpha=2/3} &= (1+\alpha+(1+\alpha^2-\alpha)^{1/2})/3|_{\alpha=2/3} \\ &= (5+\sqrt{7})/9 \\ &\cong 0.8495279. \end{aligned}$$

From (4.5), since $\sigma/\gamma = 1/40$, we have

$$x_{1/40,g}(\tfrac{2}{3}) = (\tfrac{5}{3}+(\tfrac{1}{90}-8G(\cosh\sqrt{g}-1))^{1/2})/2.$$

When $g = 0$,

$$x_{1/40,0}(\tfrac{2}{3}) > 0.88603 > x_2|_{\alpha=2/3};$$

hence we can choose g small enough such that

$$x_{1/40,g}(\tfrac{2}{3}) > x_2|_{\alpha=2/3}.$$

That means there exists a $g^* > 0$ such that the above inequality holds whenever $0 \leqq g < g^*$. Let

$$C_g(\tfrac{1}{40}, \tfrac{2}{3}) = x_{1/40,g}(\tfrac{2}{3});$$

then

$$C_g(\tfrac{1}{40}, \tfrac{2}{3}) \in (x_2|_{\alpha=2/3}, 1),$$

and

$$Q(x) = C_g(\tfrac{1}{40}, \tfrac{2}{3})(\sinh\sqrt{g}(x-j)+\sinh\sqrt{g}(j+1-x))/\sinh\sqrt{g}, \qquad j \leqq x \leqq j+1$$

is a solution of (4.1)-(4.2).

Since

$$\begin{aligned} \min_{j\leqq x\leqq j+1} Q(x) &= Q\left(j+\frac{1}{2}\right) \\ &= C_g\left(\frac{1}{40}, \frac{2}{3}\right)\left(\frac{1}{\cosh\sqrt{g/2}}\right), \end{aligned}$$

there exists a g_* satisfying $g^* \geqq g_* > 0$ such that

$$\min_{j\leqq x\leqq j+1} Q(x) > x_2|_{\alpha=2/3} \quad \text{whenever } 0 \leqq g \leqq g_*.$$

Since

$$\max_{j\leqq x\leqq j+1} Q(x) = Q(j) = Q(j+1) = C_g(\tfrac{1}{40}, \tfrac{2}{3}) < 1,$$

we can conclude that

$$Q(x) \in [x_2|_{\alpha=2/3}, 1] \quad \text{for all } x \in \mathbb{R}$$

whenever $\alpha = 2/3$, $\sigma/\gamma = 1/40$ and $0 < g < g_*$.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, San Francisco, London, 1975.

[2] J. BELL, *Some threshold results for models of myelinated nerves*, Math. Biosci., 54 (1980), pp. 181–190.

[3] ———, *Parameter dependence of conduction speed for a diffusive model of myelinated axon*, IMA J. Math. Appl. Med. Biol., 3 (1986), pp. 289–300.

[4] J. BELL AND C. COSNER, *Threshold conditions for a diffusive model of a myelinated axon*, J. Math. Biol., 18 (1983), pp. 289–300.

[5] C. COSNER, *Existence of global solutions to a model of a myelinated nerve axon*, SIAM J. Math. Anal., 18 (1987), pp. 703–710.

[6] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[7] S. P. HASTINGS, *Some Mathematical Problems Arising in Neurobiology*, CIME Lecture Notes, M. Iannelli, ed., 1980.

[8] T. KATO, *Perturbation Theory for Linear Operators*, Die Grundehren Der Mathematischen Wissenschaften in Einzeldarstellungen Vol. 132, Springer-Verlag, New York, 1966.

[9] J. RINZEL, *Integration and Propagation of Neuroelectric Signals*, in Studies in Mathematical Biology, S. A. Levin, ed., The Mathematical Association of America Studies in Mathematics, Vol. 15, Washington, DC, 1978.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 582-588, May 1989

006

THE CAUCHY PROBLEM FOR THE KORTEWEG-DE VRIES EQUATION WITH MEASURES AS INITIAL DATA*

YOSHIO TSUTSUMI†

Abstract. The Cauchy problem of the Korteweg-de Vries equation is considered with measures as initial data. Global weak solutions are constructed for any bounded positive Radon measure on $\mathbb{R}$.

Key words. Korteweg-de Vries equations, Miura transform

AMS(MOS) subject classification. 35Q20

1. Introduction and a main result. In this paper we consider the Cauchy problem for the Korteweg-de Vries equation

$$u_t + u_{xxx} - 6uu_x = 0, \quad t > 0, \quad x \in \mathbb{R}, \tag{1.1}$$

$$u(0, x) = \mu(x), \tag{1.2}$$

where $\mu(x)$ is a bounded positive Radon measure on $\mathbb{R}$.

Many papers cover the global existence of solutions for (1.1)-(1.2) (see, e.g., [1], [2], [4]-[8], [10]-[13], [17]-[18], [21], and [24]-[29]). Roughly speaking, the methods used to solve (1.1)-(1.2) can be divided into two categories. One is the inverse scattering method (see, e.g., [4]-[7], [10], [14], [17], [21], [24], and [25]) and the other is the energy (or L^2-theory) method (see, e.g., [1], [2], [11]-[13], [18], [22], and [26]-[29]). Recently, the smoothing property of solutions for (1.1)-(1.2) has attracted many mathematicians and (1.1)-(1.2) has been solved for irregular initial data. In [10], by the inverse scattering method, Kappeler shows that if the initial datum $\mu(x)$ is a real measure defined on the Borel sets on $\mathbb{R}$ and satisfies the decay condition at infinity

$$\int_{-\infty}^{\infty} (1+|x|)^N \, d|\mu|(x) < \infty, \qquad N \geqq 3, \tag{1.3}$$

then (1.1)-(1.2) has a global classical solution (see also Murray [5] and Sacks [21]). Here $|\mu|$ denotes the absolute variation of μ. On the other hand, by the energy method Kato [12] and Kruzhkov and Faminskii [13] showed that if the initial datum is in $L^2(\mathbb{R})$, then (1.1)-(1.2) has a global solution. In this paper we will prove the following theorem.

THEOREM 1.1. *Let $\mu(x)$ be a positive Radon measure on $\mathbb{R}$ such that*

$$\int_{-\infty}^{+\infty} d\mu(x) < \infty. \tag{1.4}$$

Then (1.1)-(1.2) *has a global weak solution $u(t, x)$ such that*

$$u(t, x) \in L^2((0, T) \times (-R, R)) \quad \textit{for any } T, R > 0, \tag{1.5}$$

$$u(t) \in L^\infty(0, \infty; H^{-1}(\mathbb{R})), \tag{1.6}$$

$$\iint_{\pi_0} (-u\varphi_t - u\varphi_{xxx} + 3u^2\varphi_x) \, dt \, dx = 0, \qquad \varphi \in C_0^\infty(\pi_0), \tag{1.7}$$

$$u(t) \to \mu \textit{ in } H^{-1}(\mathbb{R}) \ (\textit{a.e. } t \to +0), \tag{1.8}$$

where $\pi_0 = (0, \infty) \times \mathbb{R}$.

* Received by the editors February 11, 1988; accepted for publication August 24, 1988.

† Faculty of Integrated Arts and Sciences, Hiroshima University, Higashisenda-machi, Naka-ku, Hiroshima 730, Japan.

In Theorem 1.1 the condition of the positivity of $\mu(x)$ is a little strong, but the decay condition at infinity of the initial measure requires only the boundedness of the initial measure, that is, (1.4). If we use the inverse scattering method to solve (1.1)-(1.2), such a decay condition with weight as in (1.3) seems to be indispensable. On the other hand, the energy method cannot directly be applied to (1.1)-(1.2) with the initial datum not in $L^2(\mathbb{R})$.

Our proof is based on the energy method due to Kato [12] and Kruzhkov and Faminskii [13] and the Miura transform

$$u = v_x + v^2. \tag{1.9}$$

It is well known that if $v(t, x)$ is a smooth solution of the modified Korteweg-de Vries equation

$$v_t + v_{xxx} - 6v^2 v_x = 0, \quad t > 0, \quad x \in \mathbb{R}, \tag{1.10}$$

then the function $u(t, x)$ given by (1.9) is a solution of (1.1) (see, e.g., Miura [17]). Accordingly, we can expect that if for the initial measure μ a solution $v_0(x) \in L^2(\mathbb{R})$ exists that satisfies the equation

$$v_{0x} + v_0^2 = \mu \quad \text{in } H^{-1}(\mathbb{R}), \tag{1.11}$$

then (1.9) transforms the solution of (1.10) with $v(0) = v_0$ into the solution of the original problem (1.1)-(1.2).

Recently, many mathematicians have treated the existence problem of solutions for the related nonlinear evolution equations with measures as initial data. For example, McKean [16], Osada and Kotani [20], and Sznitman [23] study the existence and uniqueness of solutions for the Burgers equation

$$u_t + uu_x = \nu u_{xx}, \quad x \in \mathbb{R}, \quad u(0, x) = c\delta(x),$$

where $\nu, c > 0$. In [15] Liu and Pierre study the existence, uniqueness, and asymptotic behavior of solutions for the equation with no dissipativity and no dispersivity

$$u_t + f(u)_x = 0, \quad x \in \mathbb{R}, \quad u(0, x) = \delta(x).$$

In [9] Giga, Miyakawa, and Osada show the existence of solutions for the two-dimensional Navier-Stokes equation with measures as initial vorticity. In [3] Brezis and Friedman study the existence and nonexistence of solutions for the semilinear heat equation

$$u_t - \Delta u + u^p = 0, \quad x \in \mathbb{R}^n, \quad u(0, x) = c\delta(x),$$

where $c > 0$ (see also Niwa [19]).

Our plan in this paper is as follows. In § 2 we present two lemmas needed for the proof of Theorem 1.1. In § 3 we give the proof of Theorem 1.1.

We conclude this section with some notation used in the paper. We abbreviate $L^p(\mathbb{R})$ and $H^m(\mathbb{R})$ to L^p and H^m, respectively. We often use the notation $D = \partial/\partial x$. For $T \geqq 0$ we denote the set $(T, \infty) \times \mathbb{R}$ by π_T. For a positive T and a Hilbert space H, $C_w([0, T]; H)$ denotes the set of all weakly continuous functions from $[0, T]$ to H. We put

$$\rho(t, x) = \begin{cases} A \exp[-1/\{1-(t^2+x^2)\}], & t^2+x^2 < 1, \\ 0, & t^2+x^2 \geqq 1, \end{cases} \tag{1.12}$$

where A is a positive constant such that $\iint_{\mathbb{R}^2} \rho(t, x)\, dt\, dx = 1$. For $\varepsilon > 0$ we let $\rho_\varepsilon(t, x) = \varepsilon^{-2}\rho(\varepsilon^{-1}t, \varepsilon^{-1}x)$.

2. Lemmas. In this section we give two lemmas needed for the proof of Theorem 1.1.

We first show the following lemma concerning the solvability of the Riccati equation.

LEMMA 2.1. *Assume that $\mu(x)$ is a positive Radon measure on $\mathbb{R}$ satisfying* (1.4). *Then there exists a solution $u(x)$ in $L^2 \cap L^\infty$ of the equation*

$$u_x + u^2 = \mu \quad \text{in } H^{-1} \tag{2.1}$$

such that

$$\|u\|_{L^2}^2 \leqq \int_{\mathbb{R}} d\mu(x), \qquad \|u\|_{L^\infty} \leqq 2 \int_{\mathbb{R}} d\mu(x).$$

Proof. We divide the proof into two steps.

Step 1. We first show Lemma 2.1, when $\mu(x)$ is a nonnegative and C^∞ function on $\mathbb{R}$ with compact support. Then, we may let $\operatorname{supp} \mu(x) \subset [a, b]$ for some $a, b \in \mathbb{R}$.

We consider the following initial value problem:

$$u_x + u^2 = \mu, \quad x \in \mathbb{R}, \quad u(a) = 0. \tag{2.2}$$

By the unique local solvability theorem of the ordinary differential equation, we have the unique local solution $u(x)$ in C^1 of (2.2). In addition, since $\mu(x) \geqq 0$, the comparison theorem shows that as long as the solution $u(x)$ exists, $u(x)$ is nonnegative for $x \geqq a$. Since $\mu(x) = 0$ for $x \leqq a$, the solution of (2.2) exists for $x \leqq a$ and

$$u(x) = 0, \qquad x \leqq a. \tag{2.3}$$

We next prove by contradiction

$$0 \leqq u(x) \leqq \sup_{x \geqq a} \sqrt{\mu(x)}, \qquad x \geqq a. \tag{2.4}$$

We assume that an x_0 exists such that $x_0 > a$ and $u(x_0) > \sup_{x \geqq a} \sqrt{\mu(x)}$. Then we put

$$x_1 = \inf \left\{ x_1 \geqq a;\ u(y) > \sup_{x \geqq a} \sqrt{\mu(x)} \text{ for } x_1 < y < x_0 \right\}. \tag{2.5}$$

We note that unless $\mu(x)$ identically vanishes, then $x_1 < x_0$. Formulae (2.2) give us

$$u_x = -u^2 + \mu < 0, \qquad x_1 < x \leqq x_0, \tag{2.6}$$

which implies

$$u(x_1) > u(x_0). \tag{2.7}$$

On the other hand, the definition of x_1 and the continuity of $u(x)$ give us

$$\sup_{x \geqq a} \sqrt{\mu(x)} = u(x_1) < u(x_0), \tag{2.8}$$

which contradicts (2.7). This shows (2.4).

Formulae (2.4) imply that the solution $u(x)$ of (2.2) exists for $x \geqq a$. Since $\mu(x) = 0$ for $x \geqq b$, we have

$$u(x) = u(b)\{u(b)(x-b) + 1\}^{-1}, \qquad x \geqq b. \tag{2.9}$$

Formulae (2.3) and (2.9) show

$$u \in L^2 \cap L^\infty, \tag{2.10}$$

$$u(x) \to 0 \quad (x \to \pm\infty). \tag{2.11}$$

Therefore, integrating (2.2) on $\mathbb{R}$, we have

$$\|u\|_{L^2}^2 = \int_{\mathbb{R}} \mu(y)\, dy. \tag{2.12}$$

In addition, by (2.2) and (2.12), we have

$$\|Du\|_{L^1} \leqq \|u\|_{L^2}^2 + \int_{\mathbb{R}} \mu(y)\, dy \leqq 2 \int_{\mathbb{R}} \mu(y)\, dy. \tag{2.13}$$

Integrating (2.2) on $[z, x]$, we have

$$|u(x)| \leqq |u(z)| + \int_z^x u(y)^2\, dy + \int_z^x \mu(y)\, dy. \tag{2.14}$$

Letting $z \to -\infty$ in (2.14), we obtain

$$|u(x)| \leqq \|u\|_{L^2}^2 + \int_{\mathbb{R}} \mu(y)\, dy \leqq 2 \int_{\mathbb{R}} \mu(y)\, dy, \qquad x \in \mathbb{R}. \tag{2.15}$$

Thus, $u(x)$ is the desired solution of (2.1).

Step 2. By the assumptions of $\mu(x)$, we can choose a function sequence $\{f_n(x)\} \subset C_0^\infty(\mathbb{R})$ such that $f_n \geqq 0$, $\|f_n\|_{L^1} \leqq \int_{\mathbb{R}} d\mu(x)$ and $f_n \to \mu$ in H^{-1} $(n \to \infty)$. For each f_n, Step 1 gives us the solution $u_n(x)$ of the equation

$$Du_n + u_n^2 = f_n \quad \text{in } H^{-1} \tag{2.16}$$

satisfying

$$\|u_n\|_{L^2}^2 \leqq \int_{\mathbb{R}} d\mu(x), \tag{2.17}$$

$$\|Du_n\|_{L^1} \leqq 2 \int_{\mathbb{R}} d\mu(x), \tag{2.18}$$

$$\|u_n\|_{L^\infty} \leqq 2 \int_{\mathbb{R}} d\mu(x). \tag{2.19}$$

Formulae (2.17)-(2.19) and the standard compactness argument show that a subsequence $\{u_{n_k}(x)\} \subset \{u_n(x)\}$ and a limit function $u(x)$ exist such that

$$u \in L^2 \cap L^\infty, \tag{2.20}$$

$$u_{n_k} \to u \quad \text{weakly in } L^2, \tag{2.21}$$

$$u_{n_k} \to u \quad *\text{-weakly in } L^\infty, \tag{2.22}$$

$$u_{n_k}(x) \to u(x) \quad \text{a.e. on } \mathbb{R}, \tag{2.23}$$

$$Du + u^2 = \mu \quad \text{in } H^{-1}, \tag{2.24}$$

$$\|u\|_{L^2}^2 \leqq \int_{\mathbb{R}} d\mu(x), \qquad \|u\|_{L^\infty} \leqq 2 \int_{\mathbb{R}} d\mu(x). \tag{2.25}$$

These complete the proof of Lemma 2.1. □

Remark 2.1. The solution of (2.1) is not necessarily unique. For example, when $\mu(x) = \delta(x)$, for $-\infty \leqq c \leqq -1$ all the functions $u(x; c)$ defined by

$$u(x; c) = \begin{cases} (c+1)/\{(c+1)x + c\}, & x > 0, \\ 1/(x+c), & x < 0, \end{cases} \tag{2.26}$$

are the solutions of (2.1).

We next consider the Cauchy problem of the modified Korteweg-de Vries equation

$$v_t + v_{xxx} - 6v^2 v_x = 0, \quad t > 0, \quad x \in \mathbb{R}, \tag{2.27}$$

$$v(0, x) = v_0(x), \qquad x \in \mathbb{R}. \tag{2.28}$$

We have the following result.

LEMMA 2.2. *For any $v_0 \in L^2$ there exists a solution $v(t)$ of* (2.27)-(2.28) *satisfying*

$$v(t) \in C_w([0, \infty); L^2) \cap L^2(0, T; H^1(-R, R)) \quad \textit{for any } T, R > 0, \tag{2.29}$$

$$\iint_{\pi_0} (-v\varphi_t - v\varphi_{xxx} + 2v^3\varphi_x)\, dt\, dx = 0, \qquad \varphi(t, x) \in C_0^\infty(\pi_0), \tag{2.30}$$

$$v(t) \to v_0 \quad \textit{in } L^2 \quad (t \to +0), \tag{2.31}$$

$$\|v(t)\|_{L^2} \leqq \|v_0\|_{L^2}, \qquad t \geqq 0. \tag{2.32}$$

Since the proof of Lemma 2.2 is essentially identical to that of Theorem 7.1 in [12] and Theorem 2.1 in [13], we omit the proof of this lemma.

3. Proof of Theorem 1.1. In this section we prove Theorem 1.1. For that purpose, we only have to give the following proposition.

PROPOSITION 3.1. *Let $\mu(x)$ be a Radon measure on $\mathbb{R}$ satisfying* (1.4). *We assume that for $\mu(x)$ there exists a solution $v_0 \in L^2$ of the equation*

$$Dv_0 + v_0^2 = \mu \quad \textit{in } H^{-1}. \tag{3.1}$$

By Lemma 2.2 *we can construct the solution $v(t)$ of* (2.27)-(2.28) *for the above v_0. We put*

$$u = Dv + v^2. \tag{3.2}$$

Then $u(t, x)$ is the solution of (1.1)-(1.2) *satisfying* (1.5)-(1.8).

Proof. Let $\varphi(t, x) \in C_0^\infty(\pi_0)$. We set

$$\varepsilon_0 = \operatorname{supp}\{T > 0;\ \varphi(t, x) = 0 \text{ on } [0, T] \times \mathbb{R}\}.$$

Let ε be a constant with $0 < \varepsilon < \varepsilon_0/4$ and let $\gamma = \varepsilon_0/2$. For the solution $v(t)$ of (2.27)-(2.28) we put

$$v_\varepsilon(t, x) = \rho_\varepsilon * v = \iint_{\pi_0} \rho_\varepsilon(t - s, x - y) v(s, y)\, ds\, dy, \quad (t, x) \in \pi_\gamma, \quad 0 < \varepsilon < \gamma/2. \tag{3.3}$$

We note that for each $(t, x) \in \pi_\gamma$

$$\rho_\varepsilon(t - s, x - y) \in C_0^\infty(\pi_0), \qquad 0 < \varepsilon < \gamma/2 \tag{3.4}$$

as a function of s and y and that $v_\varepsilon(t, x)$ is in $C^\infty(\pi_\gamma)$ for $0 < \varepsilon < \gamma/2$. We put $u_\varepsilon = Dv_\varepsilon + v_\varepsilon^2$. Then we have

$$\begin{aligned} \frac{\partial}{\partial t} u_\varepsilon + D^3 u_\varepsilon - 6u_\varepsilon D u_\varepsilon &= (D + 2v_\varepsilon)\left(\frac{\partial}{\partial t} v_\varepsilon + D^3 v_\varepsilon - 6\rho_\varepsilon * (v^2 Dv)\right) \\ &\quad + 6(D + 2v_\varepsilon)(\rho_\varepsilon * (v^2 Dv) - v_\varepsilon^2 D v_\varepsilon), \qquad (t, x) \in \pi_\gamma. \end{aligned} \tag{3.5}$$

By (3.4) and (2.30) we conclude that the first term at the right-hand side of (3.5) vanishes identically. By (3.5) and integration by parts we have

$$\begin{aligned}&\iint_{\pi_0}(-u_\varepsilon\varphi_t-u_\varepsilon\varphi_{xxx}+3u_\varepsilon^2\varphi_x)\,dt\,dx\\&=6\iint_{\pi_0}\{\rho_\varepsilon*(v^2Dv)-v_\varepsilon^2Dv_\varepsilon\}(-D\varphi+2v_\varepsilon\varphi)\,dt\,dx\\&=2\iint_{\pi_0}(\rho_\varepsilon*v^3-v_\varepsilon^3)(D^2\varphi-2\varphi Dv_\varepsilon-2v_\varepsilon D\varphi)\,dt\,dx.\end{aligned}\tag{3.6}$$

Since (2.29) and the Gagliardo-Nirenberg inequality (see, e.g., [30, Thm. 10.1 in Part 1]) imply that $v\in L^6_{\mathrm{loc}}(\pi_\gamma)$, we obtain

$$Dv_\varepsilon\to Dv \quad \text{in } L^2_{\mathrm{loc}}(\pi_\gamma),\tag{3.7}$$

$$v_\varepsilon^2\to v^2 \quad \text{in } L^2_{\mathrm{loc}}(\pi_\gamma),\tag{3.8}$$

$$v_\varepsilon^3\to v^3 \quad \text{in } L^2_{\mathrm{loc}}(\pi_\gamma),\tag{3.9}$$

$$v_\varepsilon^4\to v^4 \quad \text{in } L^1_{\mathrm{loc}}(\pi_\gamma),\tag{3.10}$$

as $\varepsilon\to+0$. Since $u_\varepsilon=Dv_\varepsilon+v_\varepsilon^2$ and the support of φ is compact and included in π_γ, (3.7)-(3.10) give us

$$u_\varepsilon\to u \quad \text{in } L^2_{\mathrm{loc}}(\pi_\gamma),\tag{3.11}$$

$$u_\varepsilon^2\to u^2 \quad \text{in } L^1_{\mathrm{loc}}(\pi_\gamma),\tag{3.12}$$

$$\iint_{\pi_0}(\rho_\varepsilon*v^3-v_\varepsilon^3)(D^2\varphi-\varphi Dv_\varepsilon-v_\varepsilon D\varphi)\,dt\,dx\to 0,\tag{3.13}$$

as $\varepsilon\to+0$. Therefore, letting $\varepsilon\to+0$ in (3.6), we obtain by (3.11)-(3.13)

$$\iint_{\pi_0}(-u\varphi_t-u\varphi_{xxx}+3u^2\varphi_x)\,dt\,dx=0.\tag{3.14}$$

Since φ is an arbitrary function in $C_0^\infty(\pi_0)$, (3.14) shows (1.7). Formulae (2.29) and (2.32) imply (1.5)-(1.6), and (1.8) follows directly from (2.29) and (2.31). This completes the proof of Proposition 3.1. □

Now Theorem 1.1 follows immediately from Lemmas 2.1, 2.2, and Proposition 3.1.

Proof of Theorem 1.1. Let $\mu(x)$ be a positive Radon measure on $\mathbb{R}$ satisfying (1.4). Then by Lemma 2.1 we have the solution $v_0\in L^2\cap L^\infty$ of (3.1). For the above v_0 we obtain by Lemma 2.2 the solution $v(t)$ of (2.27)-(2.28). Proposition 3.1 implies that the Miura transform (3.2) translates $v(t)$ into the solution $u(t)$ of (1.1)-(1.2) satisfying (1.5)-(1.8). This completes the proof of Theorem 1.1. □

Remark 3.1. The uniqueness of solutions satisfying (1.5)-(1.8) is an open problem, and it seems to be a very interesting one. When $\mu(x)=\delta(x)$, we have many different solutions of (3.1), as stated in Remark 2.1. For the different solutions of (3.1), there exist different solutions of (2.27)-(2.28). Does the Miura transform then translate all those different solutions of (2.27)-(2.28) into only one solution of (1.1)-(1.2) or not?

Acknowledgments. The author is grateful to Professors J. Ginibre and J. C. Saut for their kind encouragement and to Professor K. Chadan for his kind hospitality at Laboratoire de Physique Théorique et Hautes Energies, Orsay.

REFERENCES

[1] J. L. BONA AND L. R. SCOTT, *Solutions of the Korteweg-de Vries equation in fractional order Sobolev spaces*, Duke Math. J., 43 (1976), pp. 87-99.

[2] J. L. BONA AND R. SMITH, *The initial value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London A, 278 (1975), pp. 555-601.

[3] H. BREZIS AND A. FRIEDMAN, *Nonlinear parabolic equations involving measures as initial data*, J. Math. Pures Appl., 62 (1983), pp. 73-97.

[4] A. COHEN MURRAY, *Existence and regularity for solutions of the Korteweg-de Vries equation*, Arch. Rat. Mech. Anal., 71 (1979), pp. 143-175.

[5] ———, *Solutions of the Korteweg-de Vries equation from irregular data*, Duke Math. J., 45 (1978), pp. 149-181.

[6] A. COHEN, *Solutions of the Korteweg-de Vries equation*, in Nonlinear Partial Differential Equations in Engineering and Applied Science, R. Sternberg, ed., Marcel Dekker, New York, 1980.

[7] ———, *Decay and regularity in the inverse scattering problem*, J. Math. Anal. Appl., 87 (1982), pp. 395-426.

[8] C. S. GARDNER, J. M. GREENE, M. D. KRUSKAL, AND R. M. MIURA, *Korteweg-de Vries equation and generalizations. VI. Methods for exact solution*, Comm. Pure Appl. Math., 27 (1974), pp. 97-133.

[9] Y. GIGA, T. MIYAKAWA, AND H. OSADA, *The two dimensional Navier-Stokes flow with measures as initial vorticity*, preprint.

[10] T. KAPPELER, *Solutions to the Korteweg-de Vries equation with irregular initial data*, Comm. Partial Differential Equations, 11 (1986), pp. 927-945.

[11] T. KATO, *On the Korteweg-de Vries equation*, Manuscripta Math., 28 (1979), pp. 89-99.

[12] ———, *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, in Studies in Applied Mathematics, V. Guillemin, ed., Adv. in Math. Supplementary Studies, 18, Academic Press, New York, 1983, pp. 93-128.

[13] S. N. KRUZHKOV AND A. V. FAMINSKII, *Generalized solutions of the Cauchy problem for the Korteweg-de Vries equation*, Math. USSR Sbornik, 48 (1984), pp. 391-421.

[14] P. D. LAX, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467-490.

[15] T.-P. LIU AND M. PIERRE, *Source-solutions and asymptotic behavior in conservation laws*, J. Differential Equations, 51 (1984), pp. 419-441.

[16] H. P. MCKEAN JR., *Propagation of chaos for a class of nonlinear parabolic equations*, Lecture Series in Differential Equations, Session 7, Catholic University, Washington, DC, 1967.

[17] R. M. MIURA, *The Korteweg-de Vries equation: A survey of results*, SIAM Rev., 18 (1976), pp. 412-459.

[18] T. MUKASA AND R. IINO, *On the global solutions for the simplest generalized Korteweg-de Vries equation*, Math. Japon., 14 (1969), pp. 75-83.

[19] Y. NIWA, *Semilinear heat equations with measures as initial data*, preprint.

[20] H. OSADA AND S. KOTANI, *Propagation of chaos for the Burgers equation*, J. Math. Soc. Japan, 37 (1985), pp. 275-294.

[21] R. L. SACKS, *Classical solutions of the Korteweg-de Vries equation for non-smooth initial data via inverse scattering*, Comm. Partial Differential Equations, 10 (1985), pp. 29-98.

[22] J. C. SAUT AND R. TEMAM, *Remarks on the Korteweg-de Vries equation*, Israel J. Math., 24 (1976), pp. 78-87.

[23] A. S. SZNITMAN, *Propagation of chaos result for the Burgers equation*, preprint.

[24] S. TANAKA, *Korteweg-de Vries equation: Construction of solutions in terms of scattering data*, Osaka J. Math., 11 (1974), pp. 49-59.

[25] ———, *Korteweg-de Vries equation: Asymptotic behavior of solutions*, Publ. Res. Inst. Math. Sci., 10 (1975), pp. 367-379.

[26] R. TEMAM, *Sur un problème non linéaire*, J. Math. Pures Appl., 48 (1969), pp. 159-172.

[27] M. TSUTSUMI, *On global solutions of the generalized Korteweg-de Vries equation*, Publ. Res. Inst. Math. Sci., 7 (1972), pp. 329-344.

[28] M. TSUTSUMI, T. MUKASA, AND R. IINO, *On the generalized Korteweg-de Vries equation*, Proc. Japan Acad., 46 (1970), pp. 921-925.

[29] M. TSUTSUMI AND T. MUKASA, *Parabolic regularization of the generalized Korteweg-de Vries equation*, Funkcial. Ekvac., 14 (1971), pp. 89-110.

[30] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[31] A. COHEN AND T. KAPPELER, *Solutions to the Korteweg-de Vries equation with initial profile in* $L^1_1(\mathbb{R}) \cap L^1_N(\mathbb{R}^+)$, SIAM J. Math. Anal., 18 (1987), pp. 991-1025.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 589-607, May 1989

007

SOMMERFELD DIFFRACTION PROBLEMS WITH THIRD KIND BOUNDARY CONDITIONS*

F.-O. SPECK†§, R. A. HURD‡§, AND E. MEISTER†

Abstract. This paper continues earlier work on diffraction problems with first and second kind boundary conditions. New operator theoretic difficulties appear for third kind conditions corresponding to different behavior of the Fourier symbol matrix of the boundary operators at infinity. Compatibility conditions force another function space setting in order to obtain closed operators. Then well-posed problems can be solved by explicit Wiener-Hopf factorization using Khrapkov's method, which also yields the asymptotics near the origin.

Key words. diffraction problem, Wiener-Hopf operator, factorization, matrix function, Helmholtz equation

AMS(MOS) subject classifications. 47B35, 35J05, 45E10

1. Introduction. From the physical point of view, third kind conditions often make more sense than Dirichlet or Neumann conditions do [21]; for instance, think of the electromagnetic theory where impedance boundary conditions on a screen represent finite as opposed to perfect conductivity. We consider the problem $\mathcal{P}$ of finding

$$\begin{aligned} &u \in L^2(\mathbb{R}^2), \\ &u|_{\Omega^\pm} \in H^1(\Omega^\pm), \qquad \Omega^\pm: x_2 \gtrless 0, \\ &(\Delta + k^2)u = 0 \quad \text{in } \Omega^\pm, \end{aligned} \tag{1.1}$$

where $\operatorname{Im} k > 0$ holds and the Dirichlet data $u_0^\pm = u|_{x_2=\pm 0}$ and the Neumann data $u_1^\pm = \partial u/\partial x_2|_{x_2=\pm 0}$ satisfy

$$\begin{aligned} a_{11}u_0^- + a_{12}u_0^+ + a_{13}u_1^- + a_{14}u_1^+ &= h_1 \quad \text{on } \mathbb{R}_+, \\ a_{21}u_0^- + a_{22}u_0^+ + a_{23}u_1^- + a_{24}u_1^+ &= h_2 \quad \text{on } \mathbb{R}_+, \end{aligned} \tag{1.2}$$

and

$$\begin{aligned} u_0^+ - u_0^- &= 0 \quad \text{on } \mathbb{R}_-, \\ u_1^+ - u_1^- &= 0 \quad \text{on } \mathbb{R}_-. \end{aligned} \tag{1.3}$$

The constants k, $a_{jl} \in \mathbb{C}$ and the functionals $h_j \in H^{-1/2}(\mathbb{R}_+)$ are given.

The transmission conditions (1.3) may be replaced by the assumption that $(\Delta + k^2)u = 0$ also holds across the negative x_1 half-axis. $u_0^\pm \in H^{1/2} = H^{1/2}(\mathbb{R})$ is a consequence of the trace theorem and $u_1^+ \in H^{-1/2}$ makes sense for solutions of the Helmholtz equation. Those are represented by the potential ansatz [20]

$$\begin{aligned} u(x_1, x_2) &= F^{-1}_{\xi \mapsto x_1}\{e^{x_2 t(\xi)}\hat{u}_0^-(\xi)1_-(x_2) + e^{-x_2 t(\xi)}\hat{u}_0^+(\xi)1_+(x_2)\} \\ &= G\begin{pmatrix} u_0^- \\ u_0^+ \end{pmatrix}(x_1, x_2), \end{aligned} \tag{1.4}$$

* Received by the editors August 10, 1987; accepted for publication August 9, 1988.

† Fachbereich Mathematik, Technische Hochschule Darmstadt, D-6100 Darmstadt, Federal Republic of Germany.

‡ National Research Council of Canada, Division of Electrical Engineering, Ottawa, Ontario, Canada K1A 0R8.

§ The work of these authors was sponsored by the Deutsche Forschungsgemeinschaft under grant Me 261/4-1 and by a fellowship in 1986-1987.

and we refer to the notation of [20]. In particular, $t(\xi)=(\xi^2-k^2)^{1/2}$ denotes the usual square root function with branch cuts along $\xi=\pm k\pm i\tau$, $\tau\geqq 0$; $1_\pm$ are the characteristic functions of $\mathbb{R}_\pm$, and the Fourier transformation F is defined by

$$\hat{u}_0^\pm(\xi)=F_{x_1\mapsto\xi}u_0^\pm(x_1)=\int_{-\infty}^{\infty}e^{ix_1\xi}u(x_1)\,dx_1. \tag{1.5}$$

The present paper generalizes the impedance and the reactance problems [10], [11], [14] as well as the mixed type Dirichlet-Neumann problem [7], [13], [17]. The first two examples lead to "the usual square root singularity" $\nabla u\sim \text{const}/|x|^{1/2}$ at $x=0$, which is also well known from Sommerfeld's half-plane problem [12], [16]. Surprisingly, the final example is governed by $\nabla u\sim\text{const}/|x|^{3/4}$, and we find $\nabla u\sim\text{const}\,|x|^{\delta/2-1}$ with $\delta\in(0,1]$ in a somewhat wider class of problems [21]. Now which type of a singularity is usual and which one is exceptional? We will answer this question at the end, because it relates to the question of whether a certain bounded Hilbert space operator $A\in\mathcal{L}(L^2(\mathbb{R})^2)$ admits a "general Wiener-Hopf operator factorization" $A=A_-A_+$ with respect to the projector $1_+\cdot$ [4], [18], [19], into bounded ($\delta=1$) or unbounded ($0<\delta<1$) operators $A_\pm$, respectively.

While problems with first and second kind conditions [21] lead to the discussion of factoring a one-parameter family of function matrices, as was done by Daniele's method [3], we obtain here one four-parameter and one six-parameter family of normalized function matrices (apart from the decomposing systems). The corresponding two-media problems with different wave numbers in $\Omega^\pm$ can be treated by analogy to [20], [21]; their functional analytic structure is similar and a fixed point principle can be used, the details of which are not repeated here.

The key lemma for factoring the 2×2 symbol matrices $\sigma=\sigma_-\sigma_+$ into upper/lower holomorphic function matrices with algebraic behavior at infinity is known from Riemann-Hilbert problems as Khrapkov's method [8], [9]. In general it applies to matrix functions depending on $\xi\in\mathbb{R}$ of the form

$$\sigma=\mu_1R_1+\mu_2R_2 \tag{1.6}$$

with rational function matrices R_j and (factorable) scalar functions μ_j. All our Wiener-Hopf matrices are of this type, since they are rational in $t(\xi)=(\xi^2-k^2)^{1/2}$.

2. The Wiener–Hopf system. We define the boundary operators (on the whole axis) as

$$\begin{aligned} &B_\pm=F^{-1}\sigma_{B_\pm}\cdot F: H^{1/2}\times H^{1/2}\to H^{\mp 1/2}\times H^{-1/2}\\ &\sigma_{B_+}=\begin{pmatrix} a_{11}+a_{13}t & a_{12}-a_{14}t\\ a_{21}+a_{23}t & a_{22}-a_{24}t\end{pmatrix},\qquad \sigma_{B_-}=\begin{pmatrix}-1 & 1\\ -t & -t\end{pmatrix}\end{aligned} \tag{2.1}$$

where $F^{-1}t\cdot F: H^{1/2}\to H^{-1/2}$ is a bijection.

PROPOSITION 2.1. *A function u solves problem $\mathcal{P}$, if and only if it can be represented by* (1.4) *where*

$$\begin{pmatrix}f_0\\ f_1\end{pmatrix}=B_-\begin{pmatrix}u_0^-\\ u_0^+\end{pmatrix} \tag{2.2}$$

satisfies the Wiener–Hopf system

$$W\begin{pmatrix}f_0\\ f_1\end{pmatrix}=1_+\cdot B_+B_-^{-1}\begin{pmatrix}f_0\\ f_1\end{pmatrix}=\begin{pmatrix}h_1\\ h_2\end{pmatrix}. \tag{2.3}$$

W acts on the subspaces $\tilde{H}^{\pm 1/2}(\mathbb{R}_+)$ *of* $H^{\pm 1/2}$ *functionals supported on* $\mathbb{R}_+$ *as a bounded linear operator*

$$W = 1_+ \cdot F^{-1}\sigma \cdot F : \tilde{H}^{1/2}(\mathbb{R}_+) \times \tilde{H}^{-1/2}(\mathbb{R}_+) \to H^{-1/2}(\mathbb{R}_+)^2 \tag{2.4}$$

with the Fourier symbol matrix

$$\sigma = \sigma_{B_+}\sigma_{B_-}^{-1} = \sigma_{B+}\frac{1}{2t}\begin{pmatrix} -t & -1 \\ t & -1 \end{pmatrix}. \tag{2.5}$$

Proof. The representation formula (1.4) for H^1 solutions of the Helmoltz equation is known [20]. The (classical) substitution $f_0 = u_0^+ - u_0^-$, $f_1 = u_1^+ - u_1^- = F^{-1}t \cdot F(-u_0^+ - u_0^-)$ is given by the bijection B_-. The transmission conditions (1.3) yield $f_0 \in \tilde{H}^{1/2}(\mathbb{R}_+)$, $f_1 \in \tilde{H}^{-1/2}(\mathbb{R}_+)$ and (1.2), (1.4) imply (2.3). The mapping properties of W and its representation (2.5) are obvious.

Conversely, a solution of (2.3) yields (1.2), (1.3) for the substituted functions $u_0^{\pm}$ in (2.2), which can be inserted into (1.4).

For brevity we put $g_0 = u_0^+ + u_0^-$, $g_1 = u_1^+ + u_1^-$ and obtain the transmission conditions (1.2) in the form

$$\begin{aligned} &\alpha_{j1}f_0 + \alpha_{j2}g_0 + \alpha_{j3}f_1 + \alpha_{j4}g_1 = h_j \quad \text{on } \mathbb{R}_+, \quad j = 1, 2, \\ &= \frac{-a_{j1}+a_{j2}}{2}f_0 + \frac{a_{j1}+a_{j2}}{2}g_0 + \frac{-a_{j3}+a_{j4}}{2}f_1 + \frac{a_{j3}+a_{j4}}{2}g_1. \end{aligned} \tag{2.6}$$

This leads to

$$\sigma = \frac{1}{t}\begin{pmatrix} \alpha_{11}t - \alpha_{14}t^2 & -\alpha_{12} + \alpha_{13}t \\ \alpha_{21}t - \alpha_{24}t^2 & -\alpha_{22} + \alpha_{23}t \end{pmatrix}. \tag{2.7}$$

Remark 2.2. Writing σ in the form (1.6), which we will use for the (function theoretic) factoring, we obtain

$$\sigma = \begin{pmatrix} \alpha_{11} & \alpha_{13} \\ \alpha_{21} & \alpha_{23} \end{pmatrix} - \frac{1}{t}\begin{pmatrix} \alpha_{14}t^2 & \alpha_{12} \\ \alpha_{24}t^2 & \alpha_{22} \end{pmatrix} = R_1 + \mu R_2 \tag{2.8}$$

and conditions like $\det R_1 \neq 0$, etc., become most important.

On the other hand, it is known that the principal part of the symbol

$$\sigma = \sigma_{pr} + \sigma_{sm} = \begin{pmatrix} -\alpha_{14}t & \alpha_{13} \\ -\alpha_{24}t & \alpha_{23} \end{pmatrix} + \frac{1}{t}\begin{pmatrix} \alpha_{11}t & -\alpha_{12} \\ \alpha_{21}t & -\alpha_{22} \end{pmatrix} \tag{2.9}$$

is responsible for functional analytical properties of W where conditions like $\det \sigma_{pr} \neq 0$ become important. So we expect several particular cases, which must be discussed separately and start with the latter aspects.

PROPOSITION 2.3. *The operator W in* (2.4)-(2.5) *is equivalent to* (*coincides up to invertible factors with*) *the "lifted" linear bounded operator* W_0 *defined by*

$$W_0 = 1_+ \cdot F^{-1}\sigma_0 \cdot F : L^2(\mathbb{R}_+)^2 \to L^2(\mathbb{R}_+)^2,$$

$$\sigma_0 = \begin{pmatrix} t_-^{-1} & 0 \\ 0 & t_-^{-1} \end{pmatrix}\sigma\begin{pmatrix} t_+^{-1} & 0 \\ 0 & t_+ \end{pmatrix} = \begin{pmatrix} \dfrac{\alpha_{11} - \alpha_{14}t}{t} & \dfrac{-\alpha_{12} + \alpha_{13}t}{t_-^2} \\ \dfrac{\alpha_{21} - \alpha_{24}t}{t} & \dfrac{-\alpha_{22} + \alpha_{23}t}{t_-^2} \end{pmatrix} \tag{2.10}$$

with $t_{\pm}(\xi) = (\xi \pm k)^{1/2}$.

Proof. See [21, Thm. 2.1].

PROPOSITION 2.4. *For any choice of the coefficients a_{jl}, W is not Fredholm—as an operator into $H^{-1/2}(\mathbb{R}_+)^2$ (see (2.4)).*

Proof. The statement is equivalent to the assertion that W_0 is not Fredholm. But this is a direct consequence of (2.10) where

$$\det \sigma_0(-\infty) = -\det \sigma_0(+\infty) \tag{2.11}$$

since $\sigma_0 \in C(\mathbb{R})^{2\times 2}$ (see [15] or [21, Thm. 2.1]).

Therefore the function space setting must be modified, if we want well-posed problems. We restrict our considerations to nondegenerating stable symbol matrices, which are physically most important. Problem $\mathcal{P}$ and the operator W are said to be of *normal type*, if

$$\begin{aligned} &\det \sigma(\xi) \neq 0, \qquad \xi \in \mathbb{R}, \\ &[\det \sigma(\xi)]^{\pm 1} = O(|\xi|^{\pm d}), \qquad \xi \to \pm\infty \end{aligned} \tag{2.12}$$

hold with $d = 0, \pm 1$ (see (2.7)).

The case $d = -1$ corresponds to a pure Dirichlet problem, so $u_0^{\pm}$ are known on $\mathbb{R}_+$, but this basic Sommerfeld problem has been completely solved before (see [20]). So we center on the other two cases: $d = 0$, where one condition is of the first kind ($\alpha_{23} = \alpha_{24} = 0 = a_{23} = a_{24}$ without loss of generality) and the other is of second or third kind; or the case $d = 1$, where both conditions are essentially different in the principal part. It is convenient to abbreviate the coefficient matrices

$$\begin{aligned} \alpha &= \left(\begin{array}{cc:cc} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \end{array}\right) = \left(\begin{array}{c:c} \alpha_0 & \alpha_1 \end{array}\right), \\ \alpha_2 &= \begin{pmatrix} \alpha_{13} & \alpha_{14} \\ \alpha_{22} & \alpha_{21} \end{pmatrix}. \end{aligned} \tag{2.13}$$

LEMMA 2.5. *The following three assertions are equivalent:*

(i) *Problem $\mathcal{P}$ is of normal type with $d = 1$ or 0;*

(ii) *There hold* $\det \sigma_0 \neq 0$ *on* $\mathbb{R}$ *and* $(d = 1)$

$$\det \alpha_1 = -\frac{1}{2} \det \begin{pmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{pmatrix} = t^{-1} \det \sigma_{pr} \neq 0 \tag{2.14}$$

or $(d = 0)$

$$\operatorname{rank} \alpha_1 = 1, \qquad \det \alpha_2 \neq 0 \tag{2.15}$$

(and $\alpha_{23} = \alpha_{24} = 0$ without loss of generality);

(iii) *The boundary operator B_+ acts bijectively as*

$$B_+ = F^{-1}\sigma_{B_+} \cdot F : H^{1/2} \times H^{1/2} \to H^{-1/2} \times H^{1/2-d}. \tag{2.16}$$

Proof. This is a consequence of

$$\sigma^{-1} = \sigma_{B_-}\sigma_{B_+}^{-1} = \frac{t}{p(t)} \begin{pmatrix} -\alpha_{22} + \alpha_{23} t & \alpha_{12} - \alpha_{13} t \\ -\alpha_{21} + \alpha_{24} t & \alpha_{11} - \alpha_{14} t \end{pmatrix} \tag{2.17}$$

(see (2.7)), where p is polynomial of degree three or two, respectively. Thus (2.12) is equivalent to the fact that σ^{-1} does not degenerate and has maximal order at infinity. The rest of the proof is obvious.

3. The Fredholm property of W, case $d=0$. First we study the case $d=0$ characterized by (2.15) and assume $\alpha_{23}=\alpha_{24}=0$ without loss of generality. Problem $\mathcal{P}$ leads to a discussion of the Wiener-Hopf operator

$$W=1_+\cdot F^{-1}\sigma\cdot F:\tilde{H}^{1/2}(\mathbb{R}_+)\times\tilde{H}^{-1/2}(\mathbb{R}_+)\to H^{-1/2}(\mathbb{R}_+)\times H^{1/2}(\mathbb{R}_+),$$

$$\sigma=\begin{pmatrix}\alpha_{11}-\alpha_{14}t & \alpha_{13}-\alpha_{12}t^{-1}\\ \alpha_{21} & -\alpha_{22}t^{-1}\end{pmatrix} \tag{3.1}$$

(see (2.4), (2.7), (2.16)). This turns out to be completely different from the operator in the case $d=1$, but very similar to the type of operators that appear in connection with first and second kind conditions [21] and correspond to the subscase $\alpha_{11}=\alpha_{12}=0$ (i.e., dropping lower-order terms) in (3.1).

THEOREM 3.1. *The operator W defined by* (3.1) (*instead of* (2.4)-(2.5)) *is Fredholm with index zero, if and only if*

$$\det\sigma(\xi)\neq 0,\qquad \xi\in\mathbb{R},$$
$$\alpha_{14}\alpha_{22}\neq\lambda\alpha_{13}\alpha_{21},\qquad \lambda\in[0,1] \tag{3.2}$$

are satisfied. Otherwise the range of W is not closed in $H^{-1/2}(\mathbb{R}_+)\times H^{1/2}(\mathbb{R}_+)$ (unless $\sigma\equiv 0$ holds).

Proof. Lifting W on $L^2(\mathbb{R}_+)^2$ (see (2.10)), we now obtain equivalence to

$$W_0=1_+\cdot F^{-1}\sigma_0\cdot F,$$

$$\sigma_0=\begin{pmatrix}t_-^{-1} & 0\\ 0 & t_-\end{pmatrix}\sigma\begin{pmatrix}t_+^{-1} & 0\\ 0 & t_+\end{pmatrix}=\begin{pmatrix}\dfrac{\alpha_{11}-\alpha_{14}t}{t} & \dfrac{-\alpha_{12}+\alpha_{13}t}{t_-^2}\\ \alpha_{21}\dfrac{t_-}{t_+} & -\alpha_{22}\end{pmatrix}, \tag{3.3}$$

and the Fredholm property of W_0 is equivalent to [21, Thm. 2.1]

$$\det\sigma_0(\xi)\neq 0,\qquad \xi\in\mathbb{R},$$
$$\det[\mu\sigma_0(-\infty)+(1-\mu)\sigma_0(+\infty)]\neq 0,\qquad \mu\in[0,1]. \tag{3.4}$$

Since $t_-/t_+\to\pm 1$ at $\pm\infty$, the last assertion may be rewritten as $\alpha_{14}\alpha_{22}-(1-2\mu)^2\alpha_{13}\alpha_{21}\neq 0$, $\mu\in[0,1]$. The winding number of $\det\sigma_0(\xi)$ vanishes automatically, since

$$\det\sigma_0(\xi)=\frac{-\alpha_{11}\alpha_{22}+\alpha_{12}\alpha_{21}}{t(\xi)}+\alpha_{14}\alpha_{22}-\alpha_{13}\alpha_{21}$$
$$=-\frac{\det\alpha_0}{t(\xi)}-\det\alpha_2=\det\sigma(\xi) \tag{3.5}$$

is an even function of ξ. This yields $\operatorname{Ind} W=\operatorname{Ind} W_0=-\operatorname{ind}\det\sigma_0=0$.

COROLLARY 3.2. *For W in* (3.1) *to be Fredholm, it is necessary for problem $\mathcal{P}$ to be of normal type* (*compare* (3.2) *with* (2.15)).

COROLLARY 3.3. *The set of parameters α_{jl} where W is Fredholm is characterized by the following three conditions: first,*

$$\det\alpha_2=\alpha_{13}\alpha_{21}-\alpha_{14}\alpha_{22}\neq 0, \tag{3.6a}$$

$$\alpha_{14}\alpha_{22}\neq 0; \tag{3.6b}$$

second,

$$\det\alpha_0=\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21}=0\quad \textit{or} \tag{3.7a}$$

$$\det\alpha_0\neq 0,\qquad -\det\alpha_2/\det\alpha_0\notin\Gamma=\{\zeta=t^{-1}(\xi),\ \xi\in\dot{\mathbb{R}}\}; \tag{3.7b}$$

and third,

$$\alpha_{13}\alpha_{21}=0 \quad \textit{or}$$

$$\alpha_{13}\alpha_{21}\neq 0, \qquad \lambda=\frac{\alpha_{14}\alpha_{22}}{\alpha_{13}\alpha_{21}}\notin[0,1]. \tag{3.8}$$

(*The very last condition includes* (3.6).)

Remark 3.4. The special case $\alpha_{11}=\alpha_{12}=0$ of first and second kind conditions corresponds to (3.7a) (see [21, Thm. 2.1]).

Let us note how to proceed when (3.6)-(3.8) are violated.

(1) For $\alpha_{14}\alpha_{22}=0$, one or two compatibility conditions must be satisfied in order to get a well-posed problem; see the following example. It is easy to treat all these cases by analogy.

(2) $\det\alpha_2=0$ or $-\det\alpha_2/\det\alpha_0\in\Gamma$ leads to a problem of nonnormal type. We may introduce weighted spaces, as in the theory of singular integral operators [15], but those cases seem not to be of great physical importance (e.g., u_0^+ and u_1^+ are given on $\mathbb{R}_+$).

(3) If $\lambda\in(0,1)$ holds, replace H^s by $W^{p,s}$ spaces with $p\neq 2$, such that the symbol becomes p-regular [15, Chap. V] and the Wiener-Hopf operator is closed.

Example 3.5. The reactance problem [10], [14]

$$u_0^+=u_0^- \quad (=u_0) \quad \text{on } \mathbb{R}_+, \tag{3.9a}$$

$$u_1^+-u_1^-+\kappa u_0=h_1\in H^{-1/2}(\mathbb{R}_+) \tag{3.9b}$$

leads to conditions (2.6) of the form

$$f_1+\frac{\kappa}{2}g_0=h_1, \qquad f_0=0 \tag{3.10}$$

and therefore

$$\alpha=\begin{pmatrix}\alpha_0 & \vdots & \alpha_1\end{pmatrix}=\begin{pmatrix}0 & \frac{\kappa}{2} & 1 & 0\\ 1 & 0 & 0 & 0\end{pmatrix},$$

$$\alpha_2=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \qquad \sigma=\begin{pmatrix}0 & 1-\frac{\kappa}{2t}\\ 1 & 0\end{pmatrix}. \tag{3.11}$$

Thus we have a normal type problem with $d=0$ and the Wiener-Hopf system decomposes. All assertions (3.6)-(3.8) except (3.6b) (because of $\alpha_{14}=\alpha_{22}=0$) are satisfied for reasonable κ ($2/\kappa\notin\Gamma$). Consequently $W=1_+\cdot F^{-1}\sigma\cdot F$ is not Fredholm in the sense of (3.1), but it obviously maps $\tilde{H}^{1/2}(\mathbb{R}_+)\times\tilde{H}^{-1/2}(\mathbb{R}_+)$ into $\tilde{H}^{-1/2}(\mathbb{R}_+)\times\tilde{H}^{1/2}(\mathbb{R}_+)$ where it is invertible (see [14]). This means two compatibility conditions must be satisfied in order to obtain a well-posed problem. The first one, $h_1\in\tilde{H}^{-1/2}(\mathbb{R}_+)$, is obvious and the second, $h_2=u_0^+-u_0^-=0\in\tilde{H}^{1/2}(\mathbb{R}_+)$, is implicitly contained in (3.9a).

Example 3.6. The mixed Dirichlet-Neumann problem [5], [13], [17]

$$u_1^+=h_1\in H^{-1/2}(\mathbb{R}_+), \qquad u_0^-=h_2\in H^{1/2}(\mathbb{R}_+) \tag{3.12}$$

leads by way of (2.6) and (3.1) to

$$\alpha=\begin{pmatrix}0 & 0 & 1/2 & 1/2\\ -1/2 & 1/2 & 0 & 0\end{pmatrix},$$

$$\sigma=-\frac{1}{2}\begin{pmatrix}t & -1\\ 1 & t^{-1}\end{pmatrix}. \tag{3.13}$$

The conditions (3.6)-(3.8) are satisfied with $\det \alpha_0 = 0$ and $\lambda = -1$. W is Fredholm with index zero. It is even invertible in the sense of (3.1) [21] for all the above-mentioned data h_1, h_2.

4. The Fredholm property, case $d = 1$. In the case $d = 0$ we had to replace one data space $H^{-1/2}(\mathbb{R}_+)$ by $H^{1/2}(\mathbb{R}_+)$, since the order of a boundary condition was not maximum. Now another modification becomes relevant.

LEMMA 4.1. *If* $\det \alpha_1 \neq 0$ *holds, the compatibility condition*

$$\alpha_{24} h_1 - \alpha_{14} h_2 \in \tilde{H}^{-1/2}(\mathbb{R}_+) \tag{4.1}$$

(*instead of* $H^{-1/2}(\mathbb{R}_+)$) *is necessary for the problem* $\mathcal{P}$ *to be solvable.*

Proof. We simplify the Wiener-Hopf system (2.3), premultiply by α_1^{-1}, and obtain the system

$$\begin{aligned} &\tilde{W}\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = 1_+ \cdot F^{-1}\tilde{\sigma} \cdot F\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = \begin{pmatrix} \tilde{h}_1 \\ \tilde{h}_2 \end{pmatrix} = \alpha_1^{-1}\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \\ &\tilde{\sigma} = \alpha_1^{-1}\sigma = \begin{pmatrix} \tilde{\alpha}_{11} & 1 - \dfrac{\tilde{\alpha}_{12}}{t} \\ -t + \tilde{\alpha}_{21} & -\dfrac{\tilde{\alpha}_{22}}{t} \end{pmatrix} \end{aligned} \tag{4.2}$$

with the same space setting (2.4). The Wiener-Hopf equation due to the first line reads

$$\tilde{\alpha}_{11} f_0 + f_1 - \tilde{\alpha}_{12} 1_+ \cdot F^{-1} t^{-1} F f_1 = \tilde{h}_1 = \frac{\alpha_{24} h_1 - \alpha_{14} h_2}{\det \alpha_1}. \tag{4.3}$$

If f_0 and f_1 represent the Dirichlet and Neumann data jumps on $x_2 = 0$ of a solution of problem $\mathcal{P}$, then the trace theorem and embedding yield $f_1 \in \tilde{H}^{-1/2}(\mathbb{R}_+)$, $f_0 \in \tilde{H}^{1/2}(\mathbb{R}_+) \subsetneq \tilde{H}^{-1/2}(\mathbb{R}_+)$, $1_+ \cdot F^{-1} t^{-1} F f_1 \in H^{1/2}(\mathbb{R}_+) \subsetneq \tilde{H}^{-1/2}(\mathbb{R}_+)$, which implies (4.1).

THEOREM 4.2. *Let* $\det \alpha_1 \neq 0$. *Then*

$$\tilde{W} : \tilde{H}^{1/2}(\mathbb{R}_+) \times \tilde{H}^{-1/2}(\mathbb{R}_+) \to \tilde{H}^{-1/2}(\mathbb{R}_+) \times H^{-1/2}(\mathbb{R}_+) \tag{4.4}$$

defined by (4.2) *is a Fredholm operator with index zero, if and only if*

$$\det \tilde{\sigma}(\xi) = \frac{\det \sigma(\xi)}{\det \alpha_1} \neq 0, \qquad \xi \in \mathbb{R} \tag{4.5}$$

holds, i.e., if and only if problem $\mathcal{P}$ *is of normal type* ($d = 1$).

Proof. For simplicity we assume that α_1 is the unit matrix. Note that $\tilde{W}$ maps into a different space than W (see (2.4) and Proposition 2.4).

First we use an idea of [14] for replacing the tilde space on the right of (4.4), extending it to treat coupled systems. We know from the Sommerfeld (Dirichlet) problem [20] that

$$W_S = 1_+ \cdot F^{-1} t^{-1} \cdot F : \tilde{H}^{-1/2}(\mathbb{R}_+) \to H^{1/2}(\mathbb{R}_+) \tag{4.6}$$

is a bijection with

$$W_S^{-1} = F^{-1} t_+ \cdot F 1_+ \cdot F^{-1} t_- \cdot F l \tag{4.7}$$

where $l : H^{1/2}(\mathbb{R}_+) \to H^{1/2}$ denotes any extension onto the axis. So we transform the system (4.2) (dropping the tildes) by substituting

$$u_+ = W_S f_1 \in H^{1/2}(\mathbb{R}_+) \tag{4.8}$$

and applying W_S to the first equation, which becomes

$$\alpha_{11} W_S f_0 + u_+ - \alpha_{12} W_S u_+ = W_S h_1. \tag{4.9a}$$

The second equation is unaltered except by the substitution (4.8):

$$-1_+ F^{-1} t \cdot F f_0 + \alpha_{21} f_0 - \alpha_{22} u_+ = h_2. \tag{4.9b}$$

In operator notation, we obtain equivalently

$$\begin{gathered} \tilde{\tilde{W}} \begin{pmatrix} f_0 \\ u_+ \end{pmatrix} = 1_+ \cdot F^{-1} \begin{pmatrix} \dfrac{\alpha_{11}}{t} & 1 - \dfrac{\alpha_{12}}{t} \\ \alpha_{21} - t & -\alpha_{22} \end{pmatrix} \cdot F \begin{pmatrix} f_0 \\ u_+ \end{pmatrix} = \begin{pmatrix} W_S h_1 \\ h_2 \end{pmatrix}, \\ \tilde{\tilde{W}} : X = \tilde{H}^{1/2}(\mathbb{R}_+) \times H^{1/2}(\mathbb{R}_+) \to Y = H^{1/2}(\mathbb{R}_+) \times H^{-1/2}(\mathbb{R}_+). \end{gathered} \tag{4.10}$$

Now consider the extension of $\tilde{\tilde{W}}$:

$$\bar{W} : \bar{X} = \tilde{H}^{1/2}(\mathbb{R}_+) \times L^2(\mathbb{R}_+) \to \bar{Y} = L^2(\mathbb{R}_+) \times H^{-1/2}(\mathbb{R}_+), \tag{4.11}$$

which is also linear bounded (see the orders in (4.10)). Next we show the statement for $\bar{W}$ instead of $\tilde{W}$.

The lifting procedure leads to the equivalent operator

$$\begin{gathered} \bar{W}_0 : 1_+ \cdot F^{-1} \bar{\sigma}_0 \cdot F : L^2(\mathbb{R}_+)^2 \to L^2(\mathbb{R}_+)^2, \\ \bar{\sigma}_0 = \begin{pmatrix} 1 & 0 \\ 0 & t_-^{-1} \end{pmatrix} \tilde{\tilde{\sigma}} \begin{pmatrix} t_+^{-1} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \dfrac{\alpha_{11}}{t t_+} & 1 - \dfrac{\alpha_{12}}{t} \\ \dfrac{\alpha_{21}}{t} - 1 & \dfrac{-\alpha_{22}}{t_-} \end{pmatrix}, \end{gathered} \tag{4.12}$$

whose symbol matrix is in $C(\dot{\mathbb{R}})^{2\times 2}$, continuous at infinity, and we get

$$\begin{aligned} \det \bar{\sigma}_0 &= t^{-1} \det \tilde{\tilde{\sigma}} = t^{-1} \det \tilde{\sigma} \\ &= 1 - \frac{\alpha_{12} + \alpha_{21}}{t} + \frac{\alpha_{12}\alpha_{21} - \alpha_{11}\alpha_{22}}{t^2}. \end{aligned} \tag{4.13}$$

Thus the Fredholm property of $\bar{W}_0$ is equivalent to (4.5) (see [15]). Further we obtain

$$\operatorname{Ind} \bar{W}_0 = -\operatorname{ind} \det \bar{\sigma}_0 = -\arg \det \bar{\sigma}_0(\xi)\big|_{\xi=-\infty}^{+\infty} = 0 \tag{4.14}$$

since $\det \bar{\sigma}_0$ is an even function.

Finally we prove that the kernels and co-kernels of $\tilde{\tilde{W}}$ and $\bar{W}$ have the same dimensions. Because a solution of (4.10) for a given function in Y is automatically in X, we obtain $\ker \tilde{\tilde{W}} = \ker \bar{W}$. If this is finite-dimensional, a complement X_1 of $\ker \tilde{\tilde{W}}$ in X is dense in a complement $\bar{X}_1$ of $\ker \bar{W}$ in $\bar{X}$. Thus the image $\tilde{\tilde{W}} X_1 = \bar{W} X_1$ is dense in $\bar{W}\bar{X}_1$, since $\bar{W}$ is continuous. A finite-dimensional complement of $\tilde{\tilde{W}} X_1$ is a complement of $\bar{W}\bar{X}_1$.

Remark 4.3. The above result reflects the fact that $\tilde{H}^{-1/2}(\mathbb{R}_+) \subsetneq H^{-1/2}(\mathbb{R}_+)$ is dense but not closed. For normal type problems, the images of $\tilde{W}$, $\tilde{\tilde{W}}$, and $\bar{W}$ are closed but the image of W is not. So it is surprising that the corresponding (unlifted) symbols are simply connected by constant factors,

$$\bar{\sigma} = \frac{1}{\alpha_{12}} \begin{pmatrix} 1 & 0 \\ -\alpha_{22} & \alpha_{12} \end{pmatrix} \sigma \begin{pmatrix} 1 & 0 \\ -\alpha_{11} & \alpha_{12} \end{pmatrix}, \tag{4.15}$$

although the lifted symbol matrices are generalized factorable or not, respectively, because of the Fredholm criterion. This shows that function-theoretic factorization

(with just algebraic behavior at infinity) and generalized matrix factorization (related to the function spaces) are quite different—although related—topics.

COROLLLARY 4.4. *Condition* (4.5) *is violated, if and only if one of the characteristic numbers*

$$\lambda_{\pm}=\frac{\tilde{\alpha}_{12}+\tilde{\alpha}_{21}}{2}\pm\frac{1}{2}[(\tilde{\alpha}_{12}-\tilde{\alpha}_{21})^2+4\tilde{\alpha}_{11}\tilde{\alpha}_{22}]^{1/2} \tag{4.16}$$

is situated on the curve $\{\zeta=t(\xi),\ \xi\in\mathbb{R}\}$. *In the special case* $\tilde{\alpha}_{11}\tilde{\alpha}_{22}=0$ *these numbers simply read* $\lambda_+=\tilde{\alpha}_{12}$, $\lambda_-=\tilde{\alpha}_{21}$.

Example 4.5. The impedance problem [11], [14]

$$\begin{aligned} u_1^+ + ipu_0^+ &= h_1 \quad \text{on } \mathbb{R}_+, \\ u_1^- - iqu_0^- &= h_2 \quad \text{on } \mathbb{R}_+, \end{aligned} \tag{4.17}$$

yields the coefficient matrix

$$\tilde{\alpha}=\begin{pmatrix}\tilde{\alpha}_{11} & & \\ & \ddots & \\ & & \tilde{\alpha}_{24}\end{pmatrix}=\begin{pmatrix} i(p-q)/2 & i(p+q)/2 & 1 & 0 \\ i(p+q)/2 & i(p-q)/2 & 0 & 1\end{pmatrix} \tag{4.18}$$

and symbol matrices

$$\tilde{\sigma}=\begin{pmatrix} i\dfrac{p-q}{2} & 1-i\dfrac{p+q}{2t} \\ i\dfrac{p+q}{2}-t & -i\dfrac{p-q}{2t}\end{pmatrix},$$

$$\tilde{\tilde{\sigma}}=\bar{\sigma}=\begin{pmatrix} i\dfrac{p-q}{2t} & 1-i\dfrac{p+q}{2t} \\ i\dfrac{p+q}{2}-t & -i\dfrac{p-q}{2}\end{pmatrix}, \tag{4.19}$$

$$\bar{\sigma}_0=\begin{pmatrix} i\dfrac{p-q}{2tt_+} & 1-i\dfrac{p+q}{2t} \\ i\dfrac{p+q}{2t}-1 & -i\dfrac{p-q}{2t_-}\end{pmatrix}.$$

The system decomposes if and only if $p=q$ holds; the above matrices are then antidiagonal. In any case the determinant of the lifted symbol has the nice form

$$\det\bar{\sigma}_0=1-\frac{i(p+q)}{t}-\frac{pq}{t^2}=\left(1-\frac{ip}{t}\right)\left(1-\frac{iq}{t}\right) \tag{4.20}$$

and does not vanish for positive impedances p, q.

5. Explicit factorization. In this paper we always have the lifted symbol matrices $\sigma_0\in PC(\dot{\mathbb{R}})^{2\times 2}$, which are piecewise continuous with at most one jump at infinity. We briefly recall some facts from linear operator theory. It is well known [15], [21] that the Fredholm property of

$$W_0=1_+\cdot A_0=1_+\cdot F^{-1}\sigma_0\cdot F: L^2(\mathbb{R}_+)^2\to L^2(\mathbb{R}_+)^2 \tag{5.1}$$

is equivalent to the existence of a *generalized factorization*

$$\sigma_0(\xi)=\sigma_{0-}(\xi)\begin{pmatrix}\left(\frac{\xi-i}{\xi+i}\right)^{\kappa_1} & 0\\ 0 & \left(\frac{\xi-i}{\xi+i}\right)^{\kappa_2}\end{pmatrix}\sigma_{0+}(\xi) \tag{5.2}$$

with $\kappa_j\in\mathbb{Z}$ and

$$\sigma_{0\pm},\ \sigma_{0\pm}^{-1}\in L^2(\mathbb{R},\rho),\qquad \rho(\xi)=(\xi^2+1)^{-1/2} \tag{5.3}$$

and (5.2)-(5.3) are furthermore equivalent to the conditions (3.4).

The subcase of a *right canonical factorization* of $\sigma_0\in C(\dot{\mathbb{R}})^{2\times 2}$ is characterized by $\sigma_{0\pm}\in C(\dot{\mathbb{R}})^{2\times 2}$ and corresponds to a factorization of the bounded convolution operator $A_0\in\mathscr{L}(L^2(\mathbb{R})^2)$ into bounded operators $A_{0\pm}=F^{-1}\sigma_{0\pm}\cdot F$ and C according to the diagonal Fourier symbol matrix in the middle. In general $A_{0\pm}$ and $A_{0\pm}^{-1}=F^{-1}\sigma_{0\pm}^{-1}\cdot F$ are unbounded, but combinations like $A_0^{-1}1_+\cdot A_{0-}$ represent bounded operators, if the factors are taken from a generalized factorization.

A continuously invertible Wiener-Hopf operator W_0 and thus a well-posed problem $\mathscr{P}$ corresponds to a factorization (5.2), where $\kappa_1=\kappa_2=0$ holds. But the partial indices κ_j cannot be obtained from considering $\det\sigma_0(\xi)$, which only gives $\operatorname{Ind} W_0=-\kappa_1-\kappa_2=-\operatorname{ind}\det\sigma_0$. So we really need an explicit generalized factorization of the form (5.2)-(5.3) for the decision about invertibility. The inverse then reads

$$W_0^{-1}=A_{0\pm}^{-1}1_+\cdot A_{0-}^{-1}=F^{-1}\sigma_{0+}^{-1}\cdot F1_+\cdot F^{-1}\sigma_{0-}^{-1}\cdot F \tag{5.4}$$

and yields a corresponding formula for W^{-1} or $\tilde{W}^{-1}$ with immediate consequences for the asymptotics (the case $\kappa_j\neq 0$ appears here only, if we switch from L^2 to L^p theory; see [21]).

We are going to factor explicitly the symbol matrices (3.3) and (4.2) in standard form according to the normal type cases $d=0$ with six parameters and $d=1$ with four parameters, provided the assumptions (3.2) and (4.1), (4.5), respectively, for the Fredholm property are satisfied. According to situations, which are different in the operator theoretical or function theoretical sense, the representation splits up into many cases.

The first step for finding a right canonical or a generalized factorization consists in an application of the following theorem by Khrapkov [9], which is mostly equivalent to a simpler version ($a_1=1$, $c=0$) presented independently by Daniele [3], [8] and is connected to ideas of Heins [6] and Čebotarev [1].

THEOREM 5.1. *Let $G\in C(\mathbb{R})^{2\times 2}$ be a matrix function of the form*

$$G=a_1I+a_2R \tag{5.5}$$

where a_j are scalar functions and R is a polynomial matrix of commutant form

$$R(\xi)=\begin{pmatrix}-c(\xi) & a(\xi)\\ b(\xi) & c(\xi)\end{pmatrix},\qquad \xi\in\mathbb{R}(\mathbb{C}). \tag{5.6}$$

Abbreviate the determinants by

$$\begin{aligned}-\det R&=c^2+ab=g^2f,\\ \det G&=a_1^2-a_2^2g^2f\end{aligned} \tag{5.7}$$

where f, g are polynomials and $g(\xi)\neq 0$ for $\xi\in\mathbb{R}$ (we usually want f to have minimal degree, i.e., g contains all square factors apart from those that vanish on $\mathbb{R}$).

Furthermore, let $\det G \neq 0$ *on* $\mathbb{R}$ *and let*

$$\tau = \frac{1}{\sqrt{f}} \log \frac{a_1 + a_2 g\sqrt{f}}{a_1 - a_2 g\sqrt{f}} = \tau_- + \tau_+ \tag{5.8}$$

be an additive decomposition into functions $\tau_\pm$, *which are holomorphically extendable into the upper/lower half-plane* $\mathbb{C}_\pm$: $\operatorname{Im} \xi \gtrless 0$ *and continuous on* $\overline{\mathbb{C}_\pm}$, *respectively, where any consistent branches are chosen. Analogously, let*

$$\sqrt{\det G} = \gamma_- \gamma_+ \tag{5.9}$$

be a multiplicative decomposition, i.e., a (function-theoretical) factorization and also

$$\frac{1}{g} = \frac{1}{g_-} \frac{1}{g_+} \tag{5.10}$$

according to the poles, i.e., $g_\pm(\xi) \neq 0$ *in* $\overline{\mathbb{C}_\pm}$.

Then a factorization of G into lower/upper holomorphic function matrices reads

$$\begin{aligned} G &= G_- G_+, \\ G_\pm &= \gamma_\pm \left\{ \frac{g_\mp \cosh\left(\frac{1}{2}\sqrt{f}\tau_\pm\right)}{g_\pm} I + \frac{\sinh\left(\frac{1}{2}\sqrt{f}\tau_\pm\right)}{g_\pm^2 \sqrt{f}} R \right\}. \end{aligned} \tag{5.11}$$

Proof. See [9] for the choice $g = 1$; the modified result is then obvious [8].

Remark 5.2. (1) For factoring matrices of the form (1.6) $\sigma = \mu_1 R_1 + \mu_2 R_2$ we first can try to factor $R_1 = R_{1-} R_{1+}$ (or R_2) and to consider $\sigma = R_{1-}(\mu_1 I + \mu_2 R_{1-}^{-1} R_2 R_{1+}^{-1}) R_{1+}$ where the middle factor can be rewritten in the *Khrapkov canonical form* [8]

$$\begin{aligned} \mu_1 I + \mu_2 T &= \mu_1 I + \frac{\mu_2}{2}(T + \tilde{T}) + \frac{\mu_2}{2}(T - \tilde{T}) \\ &= a_1 I + a_2 R, \\ \tilde{T} &= \begin{pmatrix} t_{22} & -t_{12} \\ -t_{21} & t_{11} \end{pmatrix}. \end{aligned} \tag{5.12}$$

For our purposes the factorization of rational matrices [2], [22] is straightforward.

(2) R is of commutant form (5.6), if and only if $R^2 = -\det R \cdot I$ holds, i.e., $R^{-1} = -R/\det R$ for regular matrices.

(3) If $\gamma_\pm^{-1}$ are also upper/lower holomorphic, then the inverses of (5.11) yield a (right) factorization $G^{-1} = G_-^{-1} G_+^{-1}$ up to rational scalar factors, since $G_\pm$ commute with each other and

$$\det G_\pm = \gamma_\pm^2 g_\mp^2 / g_\pm^2. \tag{5.13}$$

(4) The Khrapkov formula (5.11) simplifies essentially, if $f \equiv 1$ holds. A factorization of

$$a_1 + a_2 g = \mu = \mu_- \mu_+, \qquad a_1 - a_2 g = \nu = \nu_- \nu_+, \tag{5.14}$$

which are assumed to be nonzero on the line due to $\det G \neq 0$, then yields the simpler formula

$$G_\pm = \frac{1}{2}\left\{ \frac{g_\mp}{g_\pm}[\mu_\pm + \nu_\pm] I + \frac{1}{g_\pm^2}[\mu_\pm - \nu_\pm] R \right\}. \tag{5.15}$$

For our symbol algebra (matrices which are rational in t) the algebraic behavior at infinity is obvious. The only difficulty is that the numerator $g_\mp$ destroys the desired holomorphic properties of $G_\pm^{-1}$. In some cases this leads to the middle factor of (5.2).

(5) In general, if the degree of f is higher than two, the factors $G_\pm$ increase exponentially. Based on the above existence results for generalized factorizations, it can be proven that the following trick of Daniele [3], [8] helps to obtain algebraic behavior. Split a polynomial matrix of the form $p_1 I + p_2 R$ into factors with the same exponential behavior as we found in $G_\pm$ using Khrapkov's method, and split the same matrix, once more elementary with algebraic behavior. Combining appropriate factors we obtain a factorization of G with algebraic growth; see § 9, for instance.

(6) Further tricks to reduce the algebraic growth degree are already known from [21]: split off diagonal matrices as shown in (5.2) or polynomial matrices with a constant determinant such that the increase of $G_\pm$ matrix elements are cancelled simultaneously; also see § 6.

(7) The symmetry observed in the first remark (exchange the roles of R_1, R_2, if both are invertible) leads to an alternate factorization, which is equivalent to the modification $G_- G_+ = (\det R)^{-1}(G_- R)(R G_+)$ of (5.11), and yields factors

$$G'_\pm = \frac{\gamma_\pm}{g_\pm^2 f_\pm}\left\{\cosh\left(\frac{1}{2}\sqrt{f}\tau_\pm\right)R + g\sqrt{f}\sinh\left(\frac{1}{2}\sqrt{f}\tau_\pm\right)I\right\}. \tag{5.16}$$

We are now going to use these ideas in a discussion of the above-mentioned class of parameter-dependent problems, which are listed in the order of growing complexity.

6. The case $d=0$, $\alpha_{13}\alpha_{21}\neq 0$. We continue considering the matrix σ in (3.1), assume (3.2), and start the factoring procedure for the case $\alpha_{13}\alpha_{21}\neq 0$. First write σ in the Khrapkov canonical form (5.5) as

$$\sigma = \begin{pmatrix} \alpha_{11} & \alpha_{13} \\ \alpha_{21} & 0 \end{pmatrix}\{a_1 I + a_2 R\},$$

$$R = \begin{pmatrix} -c & a \\ b & c \end{pmatrix} = \begin{pmatrix} -\dfrac{\alpha_{11}\alpha_{12}-\alpha_{12}\alpha_{21}}{\alpha_{13}\alpha_{21}} & -\dfrac{2\alpha_{22}}{\alpha_{21}} \\ -\dfrac{2\alpha_{14}}{\alpha_{13}}t^2 & \dfrac{\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21}}{\alpha_{13}\alpha_{21}} \end{pmatrix}, \tag{6.1}$$

$$a_1 = 1+\frac{c}{2t}, \qquad a_2 = \frac{1}{2t};$$

see (5.12). Therefore, put

$$\begin{aligned} -\det R &= c^2 + ab = c^2 + 4\lambda t^2 \\ &= \left(\frac{\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21}}{\alpha_{13}\alpha_{21}}\right)^2 + 4\frac{\alpha_{14}\alpha_{22}}{\alpha_{13}\alpha_{21}}(\xi^2 - k^2) = f; \end{aligned} \tag{6.2}$$

see (5.7) where the characteristic parameter $\lambda = \alpha_{14}\alpha_{22}/\alpha_{13}\alpha_{21} \in \mathbb{C}\backslash[0,1]$ from (3.8) appears (the case $f = 4\lambda\xi^2$ is not excluded). We look at the asymptotic behavior of (5.8):

$$\begin{aligned} a_1 \pm a_2\sqrt{f} &= 1 + \frac{c \pm \sqrt{f}}{2t} \sim 1 \pm \sqrt{\lambda}, \qquad |\xi| \to \infty, \\ \tau &= \frac{1}{\sqrt{f}}\log\frac{a_1 + a_2\sqrt{f}}{a_1 - a_2\sqrt{f}} \sim \frac{1}{2\sqrt{\lambda}\,t}\log\frac{1+\sqrt{\lambda}}{1-\sqrt{\lambda}}. \end{aligned} \tag{6.3}$$

The Wiener-Levy theorem yields that τ is a Wiener algebra element, the Fourier transform of an L^1-function. Using the Hilbert transform projectors, we obtain the

additive decomposition $\tau = \tau_- + \tau_+$ where

$$\begin{aligned}\tau_\pm(\xi) &= \frac{1}{2}\left\{\tau(\xi) \pm \frac{1}{\pi i}\int_{-\infty}^{\infty} \frac{\tau(\zeta)}{\zeta - \xi}\, d\zeta\right\} \\ &\sim \pm \frac{1}{2\pi i}\,\frac{1}{2\sqrt{\lambda}} \log\frac{1+\sqrt{\lambda}}{1-\sqrt{\lambda}}\,\frac{-2}{\xi} \log\frac{2\xi}{ik}, \qquad \xi \to \pm\infty;\end{aligned} \tag{6.4}$$

see [21, (3.8)]. This implies

$$\begin{aligned}\frac{1}{2}\sqrt{f}\tau_\pm &\sim \mp \frac{1}{2\pi i} \log\frac{1+\sqrt{\lambda}}{1-\sqrt{\lambda}} \operatorname{sgn}\xi \log|\xi|, \qquad \xi \to \pm\infty, \\ \exp\left(\frac{1}{2}\sqrt{f}\tau_\pm\right) &= \begin{cases} O(\xi^{\pm\delta/2}), & \xi \to +\infty, \\ O(|\xi|^{\pm\delta/2}), & \xi \to -\infty, \end{cases}\end{aligned} \tag{6.5}$$

where

$$\delta = \operatorname{Re}\frac{1}{i\pi}\log\frac{1+\sqrt{\lambda}}{1-\sqrt{\lambda}} \in (0, 1]. \tag{6.6}$$

It follows that we have the same operator theoretic situation as in [21] (see formulas (3.13) there), which corresponds to the subcase $\alpha_{11} = \alpha_{12} = 0$, i.e., taking into account only the principal part of the boundary operator, in which case the factorization was more explicit due to an evaluation of $\tau_\pm$ in (6.4). Referring to those considerations we mention only the main results.

COROLLARY 6.1. *Let σ be given by* (3.1), $\det\sigma(\xi) \neq 0$, $\alpha_{13}\alpha_{21} \neq 0$, $\lambda = \alpha_{14}\alpha_{22}/\alpha_{13}\alpha_{21} \notin [0, 1]$. *With the notation* (6.1)-(6.6), *we obtain a (function-theoretical) factorization of $G = a_1 I + a_2 R = G_- G_+$ into lower/upper holomorphic function matrices with algebraic behavior at infinity given by*

$$G_\pm = (1+\lambda)^{1/4}\left\{\cosh\left(\frac{1}{2}\sqrt{f}\tau_\pm\right) I + \frac{\sinh\left(\frac{1}{2}\sqrt{f}\tau_\pm\right)}{\sqrt{f}}\, R\right\} \tag{6.7}$$

from Khrapkov's Theorem 5.1 and consequently

$$\begin{aligned}&\sigma = \tilde{\sigma}_-\tilde{\sigma}_+, \quad \tilde{\sigma}_- = \begin{pmatrix} \alpha_{11} & \alpha_{13} \\ \alpha_{21} & 0 \end{pmatrix} G_-, \quad \tilde{\sigma}_+ = G_+, \\ &\sigma_0 = \begin{pmatrix} t_-^{-1} & 0 \\ 0 & t_- \end{pmatrix} \tilde{\sigma}_-\tilde{\sigma}_+ \begin{pmatrix} t_+^{-1} & 0 \\ 0 & t_+ \end{pmatrix} = \tilde{\sigma}_{0-}\tilde{\sigma}_{0+}.\end{aligned} \tag{6.8}$$

The orders of (these candidates for a generalized factorization) $\tilde{\sigma}_{0\pm}$ read (see [21, (4.4)])

$$\begin{aligned}\operatorname{ord}\tilde{\sigma}_{0-} &= \begin{pmatrix} \frac{1}{2}(\delta+1) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(\delta+1) & \frac{1}{2}(\delta-1) \end{pmatrix}, \\ \operatorname{ord}\tilde{\sigma}_{0+} &= \begin{pmatrix} \frac{1}{2}(\delta-1) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(\delta+1) & \frac{1}{2}(\delta+1) \end{pmatrix}.\end{aligned} \tag{6.9}$$

Remark 6.2. Therefore, $\tilde{\sigma}_0 = \tilde{\sigma}_{0-}\tilde{\sigma}_{0+}$ does not represent a generalized factorization of $\sigma_0 \in PC(\dot{\mathbb{R}})^{2\times 2}$ (where only orders less than $\frac{1}{2}$ are allowed (cf. [21, Lemma 4.1])). But the inverse factor matrices are also lower/upper holomorphic; see (6.7) and Remark 5.2(3).

COROLLARY 6.3. *A generalized factorization of σ_0 is obtained by putting*

$$\sigma_0 = \sigma_{0-}\sigma_{0+} = \tilde{\sigma}_{0-}\begin{pmatrix} 1 & 0 \\ \xi/\sqrt{\lambda} & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\xi/\sqrt{\lambda} & 1 \end{pmatrix}\tilde{\sigma}_{0+} \tag{6.10}$$

(*cf.* [21, 4.4]), *since the asymptotics are governed by the principal part analogue with*

$$\text{ord } \sigma_{0-} = \begin{pmatrix} \frac{1}{2}(1-\delta) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(1-\delta) & \frac{1}{2}(\delta-1) \end{pmatrix},$$
$$\text{ord } \sigma_{0+} = \begin{pmatrix} \frac{1}{2}(\delta-1) & \frac{1}{2}(\delta-1) \\ \frac{1}{2}(1-\delta) & \frac{1}{2}(1-\delta) \end{pmatrix}. \tag{6.11}$$

Thus the corresponding translation invariant operators $A_{0\pm} = F^{-1}\sigma_{0\pm} \cdot F$ *are bounded on* $L^2(\mathbb{R}^2)$, *if and only if* $\delta = 1$ *is satisfied, which corresponds to* $\lambda \in (1, \infty)$.

COROLLARY 6.4. *Under the same assumptions the following statements hold*:

(1) *The lifted Wiener-Hopf operator* $W_0 = 1_+ \cdot A_0 = 1_+ \cdot F^{-1}\sigma_0 \cdot F : L^2(\mathbb{R}_+)^2 \to L^2(\mathbb{R}_+)^2$ *is invertible as*

$$W_0^{-1} = A_{0+}^{-1}1_+ \cdot A_{0-}^{-1} \tag{6.12}$$

(*defined on a dense subspace, if* $\delta \neq 1$);

(2) *The Wiener-Hopf operator* W *in* (3.1) *is invertible as*

$$W^{-1} = F^{-1}\sigma_+^{-1} \cdot F1_+ \cdot F^{-1}\sigma_-^{-1} \cdot Fl \tag{6.13}$$

with any extension l *from* $H^{-1/2}(\mathbb{R}_+) \times H^{1/2}(\mathbb{R}_+)$ *into* $H^{-1/2} \times H^{1/2}$ (*e.g., odd in the first and even in the second place*; *see* [5, 20]), *where* (6.7) *gives*

$$\sigma_- = \begin{pmatrix} \alpha_{11} & \alpha_{13} \\ \alpha_{21} & 0 \end{pmatrix} G_- \begin{pmatrix} 1 & 0 \\ \xi/\sqrt{\lambda} & 1 \end{pmatrix},$$
$$\sigma_+ = \begin{pmatrix} 1 & 0 \\ -\xi/\sqrt{\lambda} & 1 \end{pmatrix} G_+. \tag{6.14}$$

(3) *Problem* $\mathcal{P}$ *is well posed for* $h_1 \in H^{-1/2}(\mathbb{R}_+)$ *and* $h_2 \in H^{1/2}(\mathbb{R}_+)$. *The solution is given by* (1.4), (2.2)-(2.3), (6.13)-(6.14), (6.7), (6.1)-(6.4), *which yield direct a priori estimates*

$$\|u|_{\Omega^\pm}\|_{H^1(\Omega^\pm)} \leqq \text{const} \cdot \{\|h_1\|_{H^{-1/2}(\mathbb{R}_+)} + \|h_2\|_{H^{1/2}(\mathbb{R}_+)}\}. \tag{6.15}$$

COROLLARY 6.5. *Furthermore, the asymptotics of the first derivatives of the solution at the origin is given by*

$$\nabla u(x_1, x_2) \sim \text{const} \cdot \sqrt{x_1^2 + x_2^2}^{\,\delta/2-1} \tag{6.16}$$

where $\delta = \text{Re}[(i\pi)^{-1}\log((1+\sqrt{\lambda})/(1-\sqrt{\lambda}))] \in (0, 1]$ *provided the data* h_j *are sufficiently smooth. We find a square root singularity similar to Sommerfeld's half-plane problem* [12], [16], [21], *if and only if* $\delta = 1$ *holds.*

7. The case $d = 0$, $\alpha_{13}\alpha_{21} = 0$. Here we find decoupling systems and give only some hints on how to proceed (cf. [20]). Starting with (3.1)-(3.2) and $\alpha_{21} = 0$, which implies $\alpha_{14}\alpha_{22} \neq 0$, we may apply a simple factorization rule for triangular matrices (see [8, formula (1.5)]),

$$\sigma = \begin{pmatrix} \alpha_{11} - \alpha_{14}t & \alpha_{13} - \dfrac{\alpha_{12}}{t} \\ 0 & -\dfrac{\alpha_{22}}{t} \end{pmatrix} = \sigma_-\sigma_+$$
$$= \begin{pmatrix} t_-\varphi_- & t_-\varphi_-\eta_- \\ 0 & -\dfrac{\alpha_{22}}{t_-} \end{pmatrix} \begin{pmatrix} -\alpha_{14}t_+\varphi_+ & \dfrac{\eta_+}{t_+} \\ 0 & \dfrac{1}{t_+} \end{pmatrix}, \tag{7.1}$$

$$\varphi = 1 - \frac{\alpha_{11}}{\alpha_{14} t} = \varphi_- \varphi_+, \qquad \psi = \alpha_{13} - \frac{\alpha_{12}}{t},$$

(7.2)

$$\eta = \frac{t_+ \psi}{t_- \varphi_-}, \qquad H_\pm \eta = \eta_\pm .$$

$H_\pm = F 1_\pm \cdot F^{-1}$ denote the Hilbert transform projectors (see (6.4)) and all elements are in the $\pm$ Wiener subalgebra corresponding to the indices. Note that φ does not vanish on $\mathbb{R}$ due to $\det \sigma \neq 0$; its winding number is zero, since φ is an even function, and thus $\varphi_\pm^{-1}$ are also holomorphic in $\mathbb{C}_\pm$.

Considering the orders, ord $t_\pm = 1/2$ and therefore

$$A_- A_+ = F^{-1}\sigma_- F \cdot F^{-1}\sigma_+ \cdot F : H^{1/2} \times H^{-1/2} \to L^2 \times L^2 \to H^{-1/2} \times H^{1/2} \tag{7.3}$$

as linear bounded and invertible operators (see (7.1)), it is obvious that the lifted factorization (drop $t_\pm$ in (7.1)) consists also of bounded invertible operators according to $\sigma_{0\pm}^{\pm 1} \in C(\dot{\mathbb{R}})^{2\times 2}$, and the inverse symbol matrices $\sigma_{0\pm}^{-1}$ are also holomorphic in $\mathbb{C}_\pm$.

COROLLARY 7.1. *In this case $\mathcal{P}$ is well posed for any data $h_1 \in H^{-1/2}(\mathbb{R}_+)$, $h_2 \in H^{1/2}(\mathbb{R}_+)$. W has a bounded inverse as in* (6.13) *and $\nabla u \sim \text{const} \cdot (x_1^2 + x_2^2)^{-1/2}$ holds near the origin.*

Considering finally the case $\alpha_{13} = 0$, which also implies $\alpha_{14}\alpha_{22} \neq 0$ for W to be Fredholm, we can write

$$\sigma = \begin{pmatrix} \alpha_{11} - \alpha_{14} t & -\dfrac{\alpha_{12}}{t} \\ \alpha_{21} & -\dfrac{\alpha_{22}}{t} \end{pmatrix} = \begin{pmatrix} 1 & \dfrac{\alpha_{12}}{\alpha_{22}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{11} - \dfrac{\alpha_{12}\alpha_{21}}{\alpha_{22}} - \alpha_{14} t & 0 \\ \alpha_{21} & -\dfrac{\alpha_{22}}{t} \end{pmatrix} \tag{7.4}$$

which leads to a quite similar result after factoring the triangular matrix as before in (7.1).

8. The case $d = 1$, $\alpha_{11}\alpha_{12}\alpha_{21}\alpha_{22} = 0$. We have to factor the matrix $\tilde{\tilde{\sigma}} = \bar{\sigma}$ in (4.10):

$$\bar{\sigma} = \begin{pmatrix} \dfrac{\alpha_{11}}{t} & 1 - \dfrac{\alpha_{12}}{t} \\ \alpha_{21} - t & -\alpha_{22} \end{pmatrix}$$

(8.1)

$$= \begin{pmatrix} 0 & 1 \\ \alpha_{21} & -\alpha_{22} \end{pmatrix} \left\{ I + \frac{1}{\alpha_{21} t} \begin{pmatrix} \alpha_{11}\alpha_{22} - t^2 & -\alpha_{12}\alpha_{22} \\ \alpha_{11}\alpha_{21} & -\alpha_{12}\alpha_{21} \end{pmatrix} \right\}$$

considering first the case $\alpha_{21} \neq 0$. In Khrapkov form (5.5) the term in braces reads

$$G = a_1 I + a_2 R,$$

$$a_1 = 1 + \frac{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} - t^2}{2\alpha_{21} t}, \qquad a_2 = \frac{1}{2\alpha_{21} t}, \tag{8.2}$$

$$R = \begin{pmatrix} -c & a \\ b & c \end{pmatrix} = \begin{pmatrix} -[t^2 - (\alpha_{11}\alpha_{22} + \alpha_{12}\alpha_{21})] & -2\alpha_{12}\alpha_{22} \\ 2\alpha_{11}\alpha_{21} & t^2 - (\alpha_{11}\alpha_{22} + \alpha_{12}\alpha_{21}) \end{pmatrix}.$$

Further, we put

$$\begin{aligned} -\det R &= c^2 + ab \\ &= t^4 - 2t^2(\alpha_{11}\alpha_{22} + \alpha_{12}\alpha_{21}) + (\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})^2 \\ &= g^2 f. \end{aligned} \tag{8.3}$$

This contains a square factor, if $\alpha_{11}\alpha_{12}\alpha_{21}\alpha_{22}=0$ holds, and then it is even of the form

$$-\det R=(t^2\pm\det\alpha_1)^2=g^2 \tag{8.4}$$

with $f\equiv 1$ where the Khrapkov method works most effectively (see Remark 5.2(4)), unless the subexceptional case $\xi_0^2=k^2\mp\det\alpha_1\in\mathbb{R}$ is given, since real zeros have to be put into the f (see Theorem 5.1).

However, in the case $\alpha_{11}\alpha_{12}\alpha_{22}=0$, $\bar\sigma$ decomposes, and we again easily get similar results to § 7 with a bounded operator factorization, the usual square root singularity, etc. This is not repeated here. But note that the decoupling of the system (in contrast to other cases like Example 3.5) does not imply a modification of the function space setting (4.11) and (4.10). ($H^{-1/2}(\mathbb{R}_+)$ cannot be replaced by $\tilde H^{-1/2}(\mathbb{R}_+)$.)

Another more complicated problem occurs for $\alpha_{21}=0$ and $\alpha_{11}\alpha_{12}\alpha_{22}\neq 0$ where the system does not decouple, but the constant matrix in (8.1) becomes singular. So we write

$$\begin{aligned}\bar\sigma&=\begin{pmatrix}0&1\\0&-\alpha_{22}\end{pmatrix}+\frac{1}{t}\begin{pmatrix}\alpha_{11}&-\alpha_{12}\\-t^2&0\end{pmatrix}\\&=\frac{1}{t_-}\begin{pmatrix}1&0\\0&-t_-^2\end{pmatrix}\left\{I+\frac{1}{\alpha_{12}t}\begin{pmatrix}-t^2&\alpha_{11}t_-^2\\-\alpha_{22}t_+^2&\alpha_{11}\alpha_{22}\end{pmatrix}\right\}\begin{pmatrix}\alpha_{11}&-\alpha_{12}\\t_+^2&0\end{pmatrix}\frac{1}{t_+}\end{aligned} \tag{8.5}$$

and the term in braces reads in Khrapkov form

$$\begin{aligned}&G=a_1I+a_2R,\\&a_1=1-\frac{1}{2\alpha_{12}t}(t^2-\alpha_{11}\alpha_{22}),\qquad a_2=\frac{1}{2\alpha_{12}t},\\&R=\begin{pmatrix}-c&a\\b&c\end{pmatrix}=\begin{pmatrix}-(t^2+\alpha_{11}\alpha_{22})&2\alpha_{11}t_-^2\\-2\alpha_{22}t_+^2&t^2+\alpha_{11}\alpha_{22}\end{pmatrix}.\end{aligned} \tag{8.6}$$

Now, following the factorization method of § 5, we obtain

$$-\det R=c^2+ab=(t^2-\alpha_{11}\alpha_{22})^2=g^2 \tag{8.7}$$

with $f\equiv 1$, if we first consider the case $k^2+\alpha_{11}\alpha_{22}\in\mathbb{C}\backslash[0,\infty)$. Further we have to set (see (5.14))

$$\begin{aligned}&\mu=a_1+a_2g=1,\\&\nu=a_1-a_2g=1-\frac{1}{\alpha_{12}t}(t^2-\alpha_{11}\alpha_{22})=-\frac{t}{\alpha_{12}}\left(1-\frac{\alpha_{12}}{t}-\frac{\alpha_{11}\alpha_{22}}{t^2}\right)\\&\qquad=-\frac{t}{\alpha_{12}}\,\omega=\left(-\frac{t_-\omega_-}{\alpha_{12}}\right)(t_+\omega_+)=\det G,\\&g(\xi)=\xi^2-k^2-\alpha_{11}\alpha_{22}=(\xi-\xi_0)(\xi+\xi_0)=g_-(\xi)g_+(\xi)\end{aligned} \tag{8.8}$$

with $\xi_0=\sqrt{k^2+\alpha_{11}\alpha_{22}}\in\mathbb{C}_+$. Note that the bounded factors of $\omega=\omega_-\omega_+$ are unique up to a constant, since they are Wiener algebra elements. Formula (5.15) yields the factorization $G=G_-G_+$ into lower/upper holomorphic matrix functions

$$\begin{aligned}G_-&=\frac{1}{2}\left\{\frac{\xi+\xi_0}{\xi-\xi_0}\left(1-\frac{t_-\omega_-}{\alpha_{12}}\right)I+\frac{1}{(\xi-\xi_0)^2}\left(1+\frac{t_-\omega_-}{\alpha_{12}}\right)R\right\},\\G_+&=\frac{1}{2}\left\{\frac{\xi-\xi_0}{\xi+\xi_0}(1+t_+\omega_+)I+\frac{1}{(\xi+\xi_0)^2}(1-t_+\omega_+)R\right\}.\end{aligned} \tag{8.9}$$

We conclude

$$\text{ord } G_\pm = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix} \tag{8.10}$$

according to a cancellation of two terms of order $\frac{1}{2}$ in the last place. We can see from (4.12) and (8.5) that the lifted factors are bounded, since

$$\begin{aligned} &\text{ord} \begin{pmatrix} 1 & 0 \\ 0 & 1/t_- \end{pmatrix} \frac{1}{t_-} \begin{pmatrix} 1 & 0 \\ 0 & -t_-^2 \end{pmatrix} G_- = \begin{pmatrix} 0 & -1 \\ -\frac{1}{2} & 0 \end{pmatrix}, \\ &\text{ord } G_+ \begin{pmatrix} \alpha_{11} & -\alpha_{12} \\ t_+^2 & 0 \end{pmatrix} \frac{1}{t_+} \begin{pmatrix} 1/t_+ & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} & 0 \\ 0 & -1 \end{pmatrix} \end{aligned} \tag{8.11}$$

hold. But this does not give a right canonical factorization, because the factors $\xi \pm \xi_0$ in $G_\mp$ destroy the desired holomorphic properties of the inverses.

Choosing the value of the factor $\omega_+(\xi)$ at $\xi = \xi_0$ such that $1 - t_+\omega_+$ vanishes at $\xi = \xi_0$, we may split off a scalar $(\xi - \xi_0)/(\xi + \xi_0)$ from G_+, and it follows that the remainder matrix has a holomorphic inverse in $\mathbb{C}_+$. It can be shown that this determination of $\omega = \omega_-\omega_+$ yields $1 + (t_-\omega_-/\alpha_{12}) = 0$ at $\xi = -\xi_0$, such that a factor $(\xi + \xi_0)/(\xi - \xi_0)$ can be split off in G_- and cancels the previous one.

This together with (8.11) leads to a factorization of the (unlifted) symbol matrix $\bar{\sigma} = \bar{\sigma}_-\bar{\sigma}_+$ into

$$\begin{aligned} \bar{\sigma}_- &= \frac{1}{2t_-} \begin{pmatrix} 1 & 0 \\ 0 & -t_-^2 \end{pmatrix} \left\{ \left(1 - \frac{t_-\omega_-}{\alpha_{12}} \right) I + \frac{1}{\xi^2 - \xi_0^2} \left(1 + \frac{t_-\omega_-}{\alpha_{12}} \right) R \right\}, \\ \bar{\sigma}_+ &= \frac{1}{2t_+} \left\{ (1 + t_+\omega_+) I + \frac{1}{\xi^2 - \xi_0^2} (1 - t_+\omega_+) R \right\} \begin{pmatrix} \alpha_{11} & -\alpha_{12} \\ t_+^2 & 0 \end{pmatrix} \end{aligned} \tag{8.12}$$

(with $t_\pm^2 = \xi \pm k$) and the usual situation of a bounded operator factorization and the square root singularity.

The remaining exceptional case in this section is $\alpha_{21} = 0$, $\alpha_{11}\alpha_{12}\alpha_{22} \neq 0$, $\xi_0^2 = k^2 + \alpha_{11}\alpha_{22} \in [0, \infty)$. Here we have two real double zeros in (8.7). We can use the Khrapkov factorization formulas (5.11) with $g_\pm = 1$ and

$$\sqrt{f} = t^2 - \alpha_{11}\alpha_{22} = \xi^2 - k^2 - \alpha_{11}\alpha_{22} = \xi^2 - \xi_0^2. \tag{8.13}$$

The idea is [8] to compensate for the exponential behavior of $G_\pm$ at infinity by factoring a polynomial matrix of the form

$$Q = b_1 I + b_2 R \tag{8.14}$$

with rational coefficients b_j such that the factors have the inverse asymptotics.

The following more important case uses a similar argument, so we refer to analogy here.

9. The case $d = 1$, $\alpha_{11}\alpha_{12}\alpha_{21}\alpha_{22} \neq 0$. Now we continue from (8.3) writing $-\det R = f$, since it contains no square factor except the factor ξ^2 in the special case of

$$k^4 + 2k^2\lambda_2 + \lambda_1^2 = 0, \tag{9.1}$$

where we abbreviate

$$\begin{aligned} \lambda_1 &= \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} = \det \alpha_1, \\ \lambda_2 &= \alpha_{11}\alpha_{22} + \alpha_{12}\alpha_{21}. \end{aligned} \tag{9.2}$$

In order to apply the Khrapkov formulas (5.11) to (8.2), we compute

$$a_1 \pm a_2\sqrt{f} = 1 + \frac{1}{2\alpha_{21} t}[\lambda_1 - t^2 \pm (t^4 - 2\lambda_2 t^2 + \lambda_1^2)^{1/2}] \tag{9.3}$$

with $\lambda_1 \neq \pm\lambda_2$ according to the headline assumption. This yields

$$a_1 + a_2\sqrt{f} \sim 1, \quad a_1 - a_2\sqrt{f} \sim -\frac{t}{\alpha_{21}}, \quad \text{as } \xi \to \pm\infty,$$

$$\det G = a_1^2 - a_2^2 f \sim -\frac{t}{\alpha_{21}}, \tag{9.4}$$

$$\sqrt{\det G} = \gamma_- \gamma_+ \sim \sqrt{\frac{-1}{\alpha_{21}}}\, t_-^{1/2} t_+^{1/2}$$

due to a factorization in the Wiener algebra. Further,

$$\tau = \frac{1}{\sqrt{f}} \log \frac{a_1 + a_2\sqrt{f}}{a_1 - a_2\sqrt{f}} \sim \frac{1}{t^2} \log \frac{-\alpha_{21}}{t} = \tau_- + \tau_+ \tag{9.5}$$

holds, where

$$\begin{aligned} \tau_\pm(\xi) &= \frac{1}{2}\left\{ \tau(\xi) \pm \frac{1}{\pi i} \int_{-\infty}^{\infty} \frac{\tau(\zeta)}{\zeta - \xi}\, d\zeta \right\} \\ &= \frac{c_1}{\xi} + O\left(\frac{1}{\xi^2} \log |\xi| \right) \end{aligned} \tag{9.6}$$

with a constant c_1 (cf. [23]).

From the Khrapkov formulas (5.11) (with $g_\pm = 1$), we see that the factors

$$\begin{aligned} G_\pm &= \gamma_\pm \left\{ \cosh\left(\tfrac{1}{2}\sqrt{f}\tau_\pm\right) I + \frac{\sinh\left(\tfrac{1}{2}\sqrt{f}\tau_\pm\right)}{\sqrt{f}} R \right\} \\ &\sim \gamma_\pm \exp\left(\tfrac{1}{2} \operatorname{Re} c_1 |\xi|\right) \left\{ I + \frac{1}{\xi^2} R \right\} \end{aligned} \tag{9.7}$$

increase exponentially in general.

PROPOSITION 9.1. *There exists a polynomial matrix $G' = a_1' I + a_2' R$ with Khrapkov factors that behaves at infinity like $G_\pm$ up to the scalar factors $\gamma_\pm$, i.e., there holds the same formula* (9.6) *for $\tau_\pm'$.*

Proof. We consider the ansatz

$$G' = (C - t^2) I + R \tag{9.8}$$

with a constant C. For factoring we have

$$\begin{aligned} \tau' &= \frac{1}{\sqrt{f}} \log \frac{a_1' + a_2'\sqrt{f}}{a_1' - a_2'\sqrt{f}} = \frac{1}{\sqrt{f}} \log \frac{C - t^2 + \sqrt{f}}{C - t^2 - \sqrt{f}} \\ &\sim \frac{1}{t^2} \log \frac{\lambda_2 - C}{2t^2}. \end{aligned} \tag{9.9}$$

Abelian theorems for the Hilbert transformation [23] lead to (9.6), again provided C is chosen suitably (which involves elliptic integrals).

THEOREM 9.2. *In the case $d = 1$, $\alpha_{11}\alpha_{12}\alpha_{21}\alpha_{22} \neq 0$, a factorization of G in* (8.2) *with algebraic behavior at infinity is given by*

$$G = G' G_- G_-'^{-1} G_+'^{-1} G_+. \tag{9.10}$$

(*The polynomial matrix can also be put at another place between the factors.*)

Proof. Matrices of the form $G = a_1 I + a_2 R$ with a fixed commutant R form a commutative algebra. In particular, $G^{-1} = (\det G)^{-1}(a_1 I - a_2 R)$ holds (if $\det G \neq 0$) and the Khrapkov factorization of GG'^{-1} coincides with the product of the two single factorizations after a rearrangement of the factors. Therefore, (9.6) and the preceding result yield a cancelation of the exponential increase, and the stated algebraic behavior at infinity holds.

Remark 9.3. Unfortunately, the explicit representation of this factorization is quite complicated in general. So we do not work out the procedure of constructing a generalized factorization of $\bar{\sigma}$. But the main questions are already answered. First, we may directly use the classical Wiener-Hopf procedure for applications starting with (9.10) to get an explicit solution. Second, the type of the singularity is known to be the same as for the principal part problem; for instance, set $\alpha_{12} = 0$, and see § 8.

REFERENCES

[1] G. N. ČEBOTAREV, *On the closed form solution of the Riemann boundary value problem for a system of n pairs of functions*, Uchen. Zap. Kazan. Gos. Univ., 116 (1956), pp. 31-58.

[2] K. CLANCEY AND I. GOHBERG, *Factorization of Matrix Functions and Singular Integral Operators*, Birkhäuser, Basel, 1981.

[3] V. G. DANIELE, *On the factorization of Wiener-Hopf matrices in problems solvable with Hurd's method*, IEEE Trans. Antennas and Propagation, AP-26 (1978), pp. 614-616.

[4] A. DEVINATZ AND M. SHINBROT, *General Wiener-Hopf operators*, Trans. Amer. Math. Soc., 145 (1969), pp. 467-494.

[5] G. I. ÈSKIN, *Boundary Value Problems for Elliptic Pseudo-differential Equations*, American Mathematical Society, Providence, 1981. (In Russian, 1973.)

[6] A. E. HEINS, *Systems of Wiener-Hopf integral equations and their application to some boundary value problems in electromagnetic theory*, in Proc. Sympos. Appl. Math. 2 (1950), pp. 76-81.

[7] ———, *The Sommerfeld half-plane problem revisited*, II. *The factoring of a matrix of analytic functions*, Math. Methods Appl. Sci., 5 (1983), pp. 14-21.

[8] R. A. HURD, *The explicit factorization of* 2×2 *Wiener-Hopf matrices*, Preprint 1040, Fachbereich Mathematik, TH Darmstadt, 1987.

[9] A. A. KHRAPKOV, *Certain cases of the elastic equilibrium of an infinite wedge with a non-symmetric notch at the vertex, subjected to concentrated force*, Prikl. Mat. Mekch., 35 (1971), pp. 625-637.

[10] I. J. LAHAIE, *Function-theoretic techniques for the electromagnetic scattering by a resistive wedge*, Radiation Laboratory Technical Report 2, Univ. Michigan, Ann Arbor, 1981.

[11] E. LÜNEBURG AND R. A. HURD, *On the diffraction problem of a half-plane with different face impedances*, Canad. J. Phys., 62 (1984), pp. 853-860.

[12] E. MEISTER, *Randwertaufgaben der Funktionentheorie*, Teubner, Stuttgart, 1983.

[13] ———, *Some multiple-part Wiener-Hopf problems in mathematical physics*, St. Banach Center Publ., 15 (1985), pp. 359-407.

[14] E. MEISTER AND F.-O. SPECK, *Diffraction problems with impedance conditions*, Appl. Anal., 22 (1986), pp. 193-211.

[15] S. G. MIKHLIN AND S. PRÖSSDORF, *Singular Integral Operators*, Springer-Verlag, Berlin, 1987. (In German, 1980.)

[16] B. NOBLE, *Methods Based on the Wiener-Hopf Technique*, Pergamon, London, 1958.

[17] A. D. RAWLINS, *The explicit Wiener-Hopf factorization of a special matrix*, Z. Angew. Math. Mech., 61 (1981), pp. 527-528.

[18] M. SHINBROT, *On singular integral operators*, J. Math. Mech., 13 (1964), pp. 395-406.

[19] F.-O. SPECK, *General Wiener-Hopf Factorization Methods*, Pitman, London, 1985.

[20] ———, *Mixed boundary value problems of the type of Sommerfeld's half-plane problem*, Proc. Royal Soc. Edinburgh Sec. A, 104 (1986), pp. 261-277.

[21] ———, *Sommerfeld diffraction problems with first and second kind boundary conditions*, I, SIAM J. Math. Anal., 20 (1989), pp. 396-407.

[22] N. P. VEKUA, *Systems of Singular Integral Equations*, Noordhoff, Groningen, 1967.

[23] R. WONG, *Asymptotic expansion of the Hilbert transform*, SIAM J. Math. Anal., 11 (1980), pp. 92-99.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 608-623, May 1989

008

A GENERAL CONVERGENCE RESULT FOR A FUNCTIONAL RELATED TO THE THEORY OF HOMOGENIZATION*

GABRIEL NGUETSENG†

Abstract. The convergence, as $\varepsilon \downarrow 0$, of the functional $F_\varepsilon(\Psi) = \int_{\mathbb{R}^N} u_\varepsilon(x)\Psi(x, x/\varepsilon)$ associated with a given L^2 function u_ε with support in a fixed compact set is studied. The test functions $\Psi(x, y)$ are continuous on $\mathbb{R}^N \times \mathbb{R}^N$ and periodic in y. A convergence theorem is proved under the weaker assumption that u_ε remains in a bounded subset of L^2. Finally, the use of multiple-scale expansions in homogenization is justified, and a new approach is proposed for the mathematical analysis of homogenization problems.

Key words. partial differential equations, homogenization, convergence, functional, periodic

AMS(MOS) subject classifications. 35B40, 41A35

1. Introduction. The mathematical analysis of *homogenization* problems for *partial differential equations* (see [1], [9]) utilizes the functionals of the type

$$F_\varepsilon(\Psi) = \int_\Omega u_\varepsilon(x)\Psi\left(x, \frac{x}{\varepsilon}\right) dx \quad (\Omega \text{ a bounded open set in } \mathbb{R}^N).$$

The function u_ε is, say, in $L^2(\Omega)$ and is (or depends on) the solution of a partial differential equation on Ω with coefficients ε-periodic (i.e., periodic with period ε in each variable). The test function $\Psi(x, y)$ is continuous on $\bar{\Omega} \times \mathbb{R}^N$ ($\bar{\Omega}$ denotes the closure of Ω) and, for fixed x, the function $y \to \Psi(x, y)$ is periodic (with period 1 in each variable).

Let us bear in mind that for such a function, i.e., Ψ, the associated sequence $(\Psi^\varepsilon)_{\varepsilon>0}$, with $\Psi^\varepsilon(x) = \Psi(x, x/\varepsilon)$ for $x \in \Omega$, converges to the function

$$x \to \tilde{\Psi}(x) = \int_Y \Psi(x, y)\, dy \quad \text{in } L^2(\Omega)\text{-weak} \quad \text{as } \varepsilon \downarrow 0$$

(see, e.g., [1]), where $Y =]0, 1[^N$.

In view of convergence studies in the theory of homogenization two distinct situations may be considered:

(i) The sequence (u_ε) is assumed to contain a subsequence, still denoted by (u_ε) for simplicity, that converges strongly to a function u_0 in $L^2(\Omega)$ as $\varepsilon \downarrow 0$ (e.g., $u_\varepsilon \in H^1(\Omega)$, $\partial\Omega$ smooth, and (u_ε) is bounded in $H^1(\Omega)$). Hence, the corresponding sequence $(F_\varepsilon(\Psi))$ converges to the integral $\int_\Omega u_0(x)\tilde{\Psi}(x)\, dx$.

(ii) The more difficult situation, which we study here, is that in which the sequence (u_ε) only remains in a bounded subset of $L^2(\Omega)$. We may surely extract a weakly convergent subsequence, but we do not have any classical argument that allows us to pass to the limit in $F_\varepsilon(\Psi)$ for the corresponding subsequence. Indeed, for the convergence of the scalar product of two sequences in $L^2(\Omega)$, we classically need strong convergence for at least one of them.

Several aspects of this situation arise in homogenization. Let us point out two particularly interesting aspects:

(1) u_ε is some derivative of a function v_ε (i.e., $u_\varepsilon = \partial v_\varepsilon/\partial x_i$) that is the solution of a boundary value problem considered in the framework of homogenization, and the sequence (v_ε) is bounded in $H^1(\Omega)$ (see § 6). In general, this is typical of the

* Received by the editors December 17, 1986; accepted for publication (in revised form) April 26, 1988.

† Department of Mathematics, University of Yaounde, P.O.B. 812 Yaounde, Cameroon. This work was partially supported by the University of Yaounde, Cameroon.

so-called *regular* homogenization problems; that is, the class of the homogenization problems associated with a formal expansion (of the solution) of the type

$$v_\varepsilon(x) = v_0(x) + \varepsilon v_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon^2 v_2\left(x, \frac{x}{\varepsilon}\right) + \cdots, \tag{1.1}$$

where the leading term v_0, which does not depend on the local variables $y = x/\varepsilon$, "ignores" the local effects.

For the study of convergence, i.e., $\lim v_\varepsilon = v_0$ as $\varepsilon \downarrow 0$, which is one of the main objects in homogenization, we possess a method, the so-called Energy Method (see [1], [9]), that solves most of the problems of the above type. However, it does not exhibit the weak limit of the gradient $\partial v_\varepsilon/\partial x_i$, $i = 1, \cdots, N$ (that is, concretely, the local behaviour of v_ε), which is interesting from the physical point of view.

(2) u_ε is the solution of a boundary value problem whose formal analysis (in the framework of homogenization) is based on an asymptotic expansion of the type

$$u_\varepsilon(x) = u_0\left(x, \frac{x}{\varepsilon}\right) + \varepsilon u_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon^2 u_2\left(x, \frac{x}{\varepsilon}\right) + \cdots, \tag{1.2}$$

with a leading term depending on the local variables $y = x/\varepsilon$. The leading term is affected by the local effects and, consequently, there is no hope of extracting a strongly convergent subsequence from (u_ε). Here, the Energy Method becomes inoperative and, to our knowledge, there is no systematic way of proving convergence for related homogenization problems, referred to as *singular* homogenization problems (see [4], [5], [7, Chaps. 7, 8] for typical examples of this). Although we do not consider that question in this work, we believe the study of singular homogenization problems requires an appropriate approach that should be based on an extensive analysis of functionals of the type

$$F_\varepsilon(\Psi) = \int_\Omega u_\varepsilon(x)\Psi\left(x, \frac{x}{\varepsilon}\right) dx.$$

Our basic result is the proof of a convergence theorem for the functional $F_\varepsilon(\Psi) = \int_{\mathbb{R}^N} u_\varepsilon(x)\Psi(x, x/\varepsilon)\, dx$ (u_ε having its support in a fixed compact set) under the weaker hypothesis that the sequence (u_ε) remains bounded in L^2. There is no need to assume the possibility of extracting a strongly convergent subsequence.

Next, based on the above result, we give a complete justification of the use of multiple-scale asymptotic expansions (such as (1.1) or (1.2)) in the theory of homogenization: Assuming that $u_\varepsilon \in L^2(\Omega)$, with u_ε bounded in the L^2 norm, Theorem 2 gives the leading-order approximation to u_ε (in (1.2)). If u_ε lies in $H^1(\Omega)$ and is bounded in the H^1 norm, Theorem 3 gives the next-order approximation to u_ε. Theoretically, the higher-order approximations are naturally given by similar theorems provided that $u_\varepsilon \in H^2(\Omega)$ with u_ε bounded, $u_\varepsilon \in H^3(\Omega)$ with u_ε bounded, $\cdots$; however, that is quite labourious.

Finally, we propose an alternative way of proving convergence in homogenization. Our approach is carried out on a classical problem (to arrive at a correct understanding of a method, we prefer to start with a classical example). Nevertheless, we anticipate that its flexibility and its "spontaneity" make it more adaptable for unusual problems than the often very fastidious Energy Method. Indeed, the reader familiar with the so-called natural multiple-scale asymptotic method [1] will easily realize that our approach is nothing but its mathematical version. Furthermore, as we shall see in § 6, our approach exhibits the local behaviour of the solution. This is not accessible to the Energy Method, whose basic ingredient is strong convergence.

This paper is organized as follows. In § 2 we present some general notation and preliminaries. Section 3 is devoted to our basic result, the case of the whole space $\mathbb{R}^N$. In § 4 we give a more pragmatic version (in view of the theory of homogenization) of the above result, which takes into account more realistic test functions. In § 5 we prove a convergence theorem for the gradient $\partial u_\varepsilon/\partial x_i$, $i=1,\cdots,N$ (i.e., for a functional $F_\varepsilon(\Psi)$ with $\partial u_\varepsilon/\partial x_i$ in place of u_ε). In practice, such a result furnishes the next term (i.e., $u_1(x, x/\varepsilon)$ in (1.2)) in the asymptotic expansion of the solution u_ε, while the leading term is given by the theorem in § 4. Thus, the use of multiple-scale asymptotic expansions of the form (1.2) (or (1.1)) is rigorously justified in homogenization.

Finally, in § 6, we present a new approach for the mathematical analysis of homogenization problems.

We will be concerned solely with vector spaces over $\mathbb{R}$ although our result and arguments are still rigorously valid in the complex case—providing some minor modifications are made. The only measure considered in this work is the Lebesgue measure.

2. General notation and preliminaries. Let $\mathbb{R}^N (N\in\mathbb{N}, N\geqq 1)$ be the N-dimensional Euclidean space. Points in $\mathbb{R}^N$ are denoted by $x=(x_1,\cdots,x_N)$ (the global variables) or $y=(y_1,\cdots,y_N)$ (the local variables related to periodicity). The cube

$$Y=]0,1[^N=]0,1[\times\cdots\times]0,1[\ (N \text{ times})$$

is considered in the system of the local variables, with closure $\bar{Y}=[0,1]^N$.

By a Y-periodic function we mean a function on $\mathbb{R}^N$ that is periodic with period Y (i.e., with period 1 in each variable y_i).

Generally speaking, if E is a set (e.g., $\mathbb{R}^N$ or any open set in $\mathbb{R}^N$), we denote by $C(E)$ the space of continuous functions on E, by $\mathscr{K}(E)$ the space of those functions in $C(E)$ with compact supports (contained in E), and by $\mathscr{D}(E)$ the subspace of $\mathscr{K}(E)$ made up of C^∞ functions.

In connection with the periodic structure, let us introduce some specific spaces.

$C_p(\mathbb{R}^N)$ (or, for simplicity, C_p) denotes the space of functions $w\in C(\mathbb{R}^N)$, w Y-periodic.

$L^2_p(\mathbb{R}^N)$ (or L^2_p) the space of Y-periodic functions in $L^2_{\text{loc}}(\mathbb{R}^N)$, which is a Hilbert space with the norm

$$\|w\|_{L^2(Y)}=\left(\int_Y |w|^2\,dy\right)^{1/2}.$$

$\mathscr{K}(\mathbb{R}^N; L^2_p)$ the space of continuous functions on $\mathbb{R}^N$ (the Euclidean space of the variables x) with values in L^2_p and having compact supports.

$L^2(\mathbb{R}^N; L^2_p)$ the space of measurable functions $u(x,y)$ on $\mathbb{R}^N\times\mathbb{R}^N$ such that for almost all x the function $y\to u(x,y)$ belongs to L^2_p and $\int_{\mathbb{R}^N\times Y}|u(x,y)|^2\,dx\,dy<\infty$. We endow this space with the norm

$$\|u\|_{L^2(\mathbb{R}^N\times Y)}=\left[\int_{\mathbb{R}^N\times Y}|u(x,y)|^2\,dx\,dy\right]^{1/2}.$$

$L^2(\mathbb{R}^N; L^2_p)$, thus equipped, is a Hilbert space.

Finally, $\mathscr{K}(\mathbb{R}^N; C_p)$ denotes the space of continuous functions on $\mathbb{R}^N$ with values in C_p and having compact supports. We provide the vector space $\mathscr{K}(\mathbb{R}^N; C_p)$ with its natural topology: the inductive limit topology determined by the spaces $\mathscr{K}_K(\mathbb{R}^N; C_p)$ (K ranging over the compact subsets of $\mathbb{R}^N$), where

$$\mathscr{K}_K(\mathbb{R}^N; C_p)=\{\Psi\in\mathscr{K}(\mathbb{R}^N; C_p);\ \text{supp}\,\Psi\subset K\}$$

is a Banach space with the norm

$$\|\Psi\|_K = \sup_{x\in K} \|\Psi(x)\|_{L^\infty} \equiv \sup_{(x,y)\in\mathbb{R}^N\times\mathbb{R}^N} |\Psi(x,y)|$$

(note that C_p, provided with the L^∞ norm, is a Banach space).

In § 3 we will need a very useful result from Bourbaki [2, Prop. 5, p. 46]: Let $\mathscr{K}(\mathbb{R}^N)\otimes C_p$ denote the subset of $\mathscr{K}(\mathbb{R}^N; C_p)$ consisting of all functions of the form $\sum v\otimes w$ ($\otimes$ denotes the tensor product), v (respectively, w) ranging over a finite subset of $\mathscr{K}(\mathbb{R}^N)$ (respectively, C_p). Then $\mathscr{K}(\mathbb{R}^N)\otimes C_p$ is dense in $\mathscr{K}(\mathbb{R}^N; C_p)$.

Finally, for further needs, let us keep in mind the well-known result that asserts that $\mathscr{K}(\mathbb{R}^N; C_p)$ is dense in $L^2(\mathbb{R}^N; L^2_p)$.

In the sequel we will put, for simplicity,

$$\mathscr{K}_p \equiv \mathscr{K}(\mathbb{R}^N; C_p).$$

3. Basic result. A convergence theorem. In all that follows, ε, with $\varepsilon>0$, denotes a real sequence destined to tend to zero, and K_0 is a fixed compact set in $\mathbb{R}^N$ (K_0 does not depend on ε). Next, we introduce $L^2_{K_0}(\mathbb{R}^N)$, the space of all functions in $L^2(\mathbb{R}^N)$ having their (compact) supports in K_0.

3.1. Statement of the theorem. Idea of the proof.

THEOREM 1. *Let $u_\varepsilon \in L^2_{K_0}(\mathbb{R}^N)$. Suppose that there exists a constant $c>0$ such that*

$$\|u_\varepsilon\|_{L^2} \leqq c \quad \text{for any } \varepsilon. \tag{3.1}$$

Then there exist a subsequence from ε, still denoted by ε for simplicity, and a function u_0 in $L^2(\mathbb{R}^N; L^2_p)$ such that

$$\int_{\mathbb{R}^N} u_\varepsilon(x)\Psi\left(x,\frac{x}{\varepsilon}\right)dx \to \int_{\mathbb{R}^N\times Y} u_0(x,y)\Psi(x,y)\,dx\,dy \tag{3.2}$$

as $\varepsilon\downarrow 0$, for all Ψ in $\mathscr{K}_p$.

Remark 1. Instead of the cube $Y=]0,1[^N$, if we consider a parallelepiped $Y=\prod_{i=1}^N]0,a_i[(a_i>0)$, Theorem 1 remains valid provided the right-hand side of (3.2) is multiplied by $1/|Y|$ ($|Y|$ = measure of Y). □

We now give the idea and the main steps of the proof. The first step is to show that a subsequence (still denoted by ε for simplicity) can be extracted from ε such that for $w\in C_p$ the sequence $u_\varepsilon w^\varepsilon$ converges in L^2-weak as $\varepsilon\downarrow 0$, where $w^\varepsilon(x)=w(x/\varepsilon)$. Thus, given a function w in C_p, there will exist z_w in $L^2(\mathbb{R}^N)$ such that, as $\varepsilon\downarrow 0$,

$$\int_{\mathbb{R}^N} u_\varepsilon w^\varepsilon v\,dx \to \int_{\mathbb{R}^N} z_w v\,dx \quad \text{for all } v\in\mathscr{K}(\mathbb{R}^N). \tag{3.3}$$

Next, our task is to extend (3.3) (with the same subsequence ε) to all functions in $\mathscr{K}_p$ (see § 2 for the definition of $\mathscr{K}_p$). Indeed, note that the integrand on the left of (3.3) is nothing but $u_\varepsilon(x)\Psi(x,x/\varepsilon)$ with $\Psi(x,y)=v(x)w(y)$. It is then reasonable to hope that (3.3) could be generalized to all functions in $\mathscr{K}_p$. To this end, we first establish that for any Ψ in $\mathscr{K}_p$ a real number $F_0(\Psi)$ exists such that

$$\int_{\mathbb{R}^N} u_\varepsilon(x)\Psi\left(x,\frac{x}{\varepsilon}\right)dx \to F_0(\Psi). \tag{3.4}$$

This will be obtained from (3.3), because $\mathscr{K}(\mathbb{R}^N)\otimes C_p$ is dense in $\mathscr{K}_p$ (see § 2).

Finally, the last step is devoted to the characterization of the right of (3.4).

3.2. First convergence result. Our goal in this section is to obtain (3.3) for any w in C_p. To begin, let us establish two elementary (but fundamental) lemmas.

LEMMA 1. *Let K_0 be the above compact set. Fix $r>0$ and set $H=\{x\in\mathbb{R}^N;\ d(x,K_0)\leqq r\}$, where d denotes the Euclidean metric. Then for $\varepsilon<\varepsilon_0$ (ε_0 a suitable constant) there exist a natural number n (depending on ε) and a finite family $\varepsilon(\bar{Y}+k_i)$, $1\leqq i\leqq n$, with $k_i\in\mathbb{Z}^N$ ($\mathbb{Z}$ is the set of all integers) such that*

$$K_0\subset\bigcup_{i=1}^{n}\varepsilon(\bar{Y}+k_i)\subset H. \tag{3.5}$$

Proof. For arbitrarily fixed ε, we may express $\mathbb{R}^N$ (the space of the variables x) as the union of all the $\varepsilon(\bar{Y}+k)$, $k\in\mathbb{Z}^N$. Since K_0 is compact, a finite family $\varepsilon(\bar{Y}+k_i)$, $i=1,\cdots,n$, exists such that K_0 intersects each $\varepsilon(\bar{Y}+k_i)$ and K_0 is contained in their union.

Now, for each $i(1\leqq i\leqq n)$, let $x\in\varepsilon(\bar{Y}+k_i)$. Then $d(x,K_0)\leqq d(x,\varepsilon(\bar{Y}+k_i)\cap K_0)\leqq\operatorname{diam}\varepsilon(\bar{Y}+k_i)=\varepsilon\operatorname{diam}Y$ (diam denotes the diameter). Hence, by putting $\varepsilon_0=r/\operatorname{diam}Y$ it follows that for $\varepsilon<\varepsilon_0$ the union of the sets $\varepsilon(\bar{Y}+k_i)$ is contained in H, which completes the proof. □

LEMMA 2. *There exists a constant $c_0>0$ such that for $\varepsilon<\varepsilon_0$ (ε_0 is the constant in Lemma 1) we have*

$$\left|\int_{\mathbb{R}^N}u(x)w\left(\frac{x}{\varepsilon}\right)dx\right|\leqq c_0\|u\|_{L^2}\|w\|_{L^2(Y)}$$

for all u in $L^2_{K_0}(\mathbb{R}^N)$ and all w in L^2_p.

Proof. Let $u\in L^2_{K_0}(\mathbb{R}^N)$, $w\in L^2_p$. By Hölder's inequality we have

$$\left|\int_{\mathbb{R}^N}u(x)w\left(\frac{x}{\varepsilon}\right)dx\right|\leqq\|u\|_{L^2}\left[\int_{K_0}\left|w\left(\frac{x}{\varepsilon}\right)\right|^2dx\right]^{1/2}.$$

Next, by the preceding lemma, let $\varepsilon(\bar{Y}+k_i)$ $(1\leqq i\leqq n)$ be a finite family satisfying (3.5) for $\varepsilon<\varepsilon_0$. Then

$$\int_{K_0}\left|w\left(\frac{x}{\varepsilon}\right)\right|^2dx\leqq\sum_{i=1}^{n}\int_{\varepsilon(Y+k_i)}\left|w\left(\frac{x}{\varepsilon}\right)\right|^2dx.$$

By change of variable, $x=\varepsilon(y+k_i)$, and use of periodicity we have

$$\int_{\varepsilon(Y+k_i)}\left|w\left(\frac{x}{\varepsilon}\right)\right|^2dx=\varepsilon^N\int_Y|w(y)|^2\,dy.$$

It follows that

$$\int_{K_0}\left|w\left(\frac{x}{\varepsilon}\right)\right|^2dx\leqq\varepsilon^Nn\|w\|^2_{L^2(Y)}.$$

But, thanks to (3.5) we have $\varepsilon^Nn=\operatorname{meas}\bigcup_{i=1}^{n}\varepsilon(\bar{Y}+k_i)\leqq\operatorname{meas}H$ (note that n depends on ε), from which the conclusion follows (with, e.g., $c_0=(\operatorname{meas}H)^{1/2}$). □

Remark 2. For $\varepsilon<\varepsilon_0$ we have

$$\int_{K_0}\left|w\left(\frac{x}{\varepsilon}\right)\right|^2dx\leqq c_0^2\|w\|^2_{L^2(Y)}\quad\forall w\in L^2_p.$$ □

As an immediate consequence of Lemma 2, we have the following proposition, which plays an essential role throughout the rest of this section.

PROPOSITION 1. *Let $f \in L^2_{K_0}(\mathbb{R}^N)$ (f may or may not depend on ε). Then for $\varepsilon < \varepsilon_0$, a unique function $f_\varepsilon \in L^2_p$ can be assigned to f such that*

$$\int_{\mathbb{R}^N} f(x) w\left(\frac{x}{\varepsilon}\right) dx = \int_Y f_\varepsilon(y) w(y)\, dy \quad \forall w \text{ in } L^2_p,$$

$$\|f_\varepsilon\|_{L^2(Y)} \leqq c_0 \|f\|_{L^2}.$$

Remark 3. The correspondence $f \to f_\varepsilon$ defined above is linear. □

We are now in a position to prove the main result in this section. First, we must give some notation used frequently in the sequel.

Given w in L^2_p we denote by w^ε the ε-periodic function in $L^2_{\text{loc}}(\mathbb{R}^N)$ defined by

$$w^\varepsilon(x) = w\left(\frac{x}{\varepsilon}\right). \tag{3.6}$$

Also, if $\Psi \in \mathcal{K}_p$ we put

$$\Psi^\varepsilon(x) = \Psi\left(x, \frac{x}{\varepsilon}\right). \tag{3.7}$$

It is clear that $\Psi^\varepsilon \in \mathcal{K}(\mathbb{R}^N)$. Moreover, if the support of Ψ is contained in K (a compact subset of $\mathbb{R}^N$), then the support of Ψ^ε lies in K for any ε.

The aim now is to prove the following proposition.

PROPOSITION 2. *Under the assumptions of Theorem* 1, *a subsequence (still denoted by ε) can be extracted from ε such that for any w in C_p (w independent of ε), the sequence $u_\varepsilon w^\varepsilon$ converges in L^2-weak as $\varepsilon \downarrow 0$.*

Proof. (i) We begin by fixing a (nontrivial) function α in $\mathcal{D}(\mathbb{R}^N)$, α independent of ε. Next, fix x in $\mathbb{R}^N$ and consider the function $s \to f(s) = \alpha(x-s)u_\varepsilon(s)$, which belongs to $L^2_{K_0}(\mathbb{R}^N)$. By Proposition 1 there exists, for $\varepsilon < \varepsilon_0$, a unique function $y \to z_\varepsilon(x, y)$ in L^2_p such that for any w in L^2_p we have

$$\int_{\mathbb{R}^N} \alpha(x-s)u_\varepsilon(s)w^\varepsilon(s)\, ds = \int_Y z_\varepsilon(x, y)w(y)\, dy,$$

that is,

$$[(u_\varepsilon w^\varepsilon) * \alpha](x) = \int_Y z_\varepsilon(x, y)w(y)\, dy, \tag{3.8}$$

where $*$ denotes the convolution product.

Moreover, again by Proposition 1, we have

$$\|z_\varepsilon(x, \cdot)\|_{L^2(Y)} \leqq c_0 \left[\int_{\mathbb{R}^N} |\alpha(x-s)u_\varepsilon(s)|^2\, ds\right]^{1/2}. \tag{3.9}$$

Observe that the function $(u_\varepsilon w^\varepsilon) * \alpha$ lies in $\mathcal{D}(\mathbb{R}^N)$ and has its support in a compact set that does not depend on ε.

Thus, by (3.8) (valid for all x) we assign to u_ε (for $\varepsilon < \varepsilon_0$) a unique function $x \to z_\varepsilon(x)$ [i.e., $x \to z_\varepsilon(x, \cdot)$] from $\mathbb{R}^N$ to L^2_p, with (3.9).

(ii) For further needs we now study a few useful properties of the function z_ε thus constructed. To summarize, let us show that $z_\varepsilon \in L^2(\mathbb{R}^N; L^2_p)$. It suffices to check that $z_\varepsilon \in \mathcal{K}(\mathbb{R}^N; L^2_p)$ (see § 2 for notation). Clearly the function z_ε has compact support; then it remains to show continuity. For this, fix x in $\mathbb{R}^N$. Let $h \in \mathbb{R}^N$. Consider the function $s \to [\alpha(x+h-s) - \alpha(x-s)]u_\varepsilon(s)$, which lies in $L^2_{K_0}(\mathbb{R}^N)$. If we replace in (i)

the function $s \to \alpha(x-s)u_\varepsilon(s)$ by the above function, the associated analogue of $z_\varepsilon(x)$ is, according to the above process, exactly $z_\varepsilon(x+h)-z_\varepsilon(x)$ (see Remark 3). Hence,

$$\|z_\varepsilon(x+h)-z_\varepsilon(x)\|_{L^2(Y)} \leqq c_0\left[\int_{\mathbb{R}^N} |\alpha(x+h-s)-\alpha(x-s)|^2|u_\varepsilon(s)|^2\,ds\right]^{1/2},$$

which is the analogue of (3.9). Observing that the right-hand side is majorized by $cc_0 \sup_s |\alpha(x+h-s)-\alpha(x-s)|$ (c is the constant in Theorem 1) and, furthermore, α being uniformly continuous on $\mathbb{R}^N$, we deduce that $\|z_\varepsilon(x+h)-z_\varepsilon(x)\|_{L^2(Y)} \leqq c|h|$, for all $h \in \mathbb{R}^N$, which shows continuity.

Thus, $z_\varepsilon \in L^2(\mathbb{R}^N; L^2_p)$. Furthermore, by (3.9) we have

$$\|z_\varepsilon\|_{L^2(\mathbb{R}^N \times Y)} \leqq c \ (c>0) \quad \forall \varepsilon < \varepsilon_0 \tag{3.10}$$

(where the constant c does not depend on ε).

(iii) Finally, by (3.10) we can extract a subsequence from ε, still denoted by ε, such that $z_\varepsilon \to z$ in $L^2(\mathbb{R}^N; L^2_p)$-weak as $\varepsilon \downarrow 0$. Therefore, for each $v \in \mathscr{K}(\mathbb{R}^N)$ and each $w \in L^2_p$ we have

$$\int_{\mathbb{R}^N \times Y} z_\varepsilon(x,y)w(y)v(x)\,dx\,dy \to \int_{\mathbb{R}^N \times Y} z(x,y)w(y)v(x)\,dx\,dy,$$

so that, using (3.8) combined with Fubini's theorem, we have

$$\int_{\mathbb{R}^N} [(u_\varepsilon w^\varepsilon) * \alpha](x)v(x)\,dx \to \int_{\mathbb{R}^N \times Y} z(x,y)w(y)v(x)\,dx\,dy. \tag{3.11}$$

From now on, ε denotes exclusively the subsequence extracted above. By (3.11) we finally show that for each w in C_p, the sequence $u_\varepsilon w^\varepsilon$ converges weakly in $L^2(\mathbb{R}^N)$ as $\varepsilon \downarrow 0$ (that is, ε is the desired subsequence in Proposition 2). For this purpose, let w be arbitrarily fixed in C_p. Since $w^\varepsilon \in L^\infty$, we have $u_\varepsilon w^\varepsilon \in L^2$. Furthermore, we evidently have $\|u_\varepsilon w^\varepsilon\|_{L^2} \leqq c$ ($c>0$), for all ε. Therefore, we can extract ε' from ε such that

$$u_{\varepsilon'} w^{\varepsilon'} \to z_w \quad \text{in } L^2\text{-weak} \quad \text{as } \varepsilon' \downarrow 0, \tag{3.12}$$

so that, the transformation $v \to v * \alpha$ being continuous from L^2 into itself,

$$\int_{\mathbb{R}^N} [(u_{\varepsilon'} w^{\varepsilon'}) * \alpha]v\,dx \to \int_{\mathbb{R}^N} (z_w * \alpha)v\,dx$$

for all $v \in \mathscr{K}(\mathbb{R}^N)$. By comparison with (3.11) we necessarily have

$$(z_w * \alpha)(x) = \int_Y z(x,y)w(y)\,dy \quad \text{a.e. in } \mathbb{R}^N. \tag{3.13}$$

Now, since w is the same function as in (3.12), let ε'' be another subsequence from ε such that $u_{\varepsilon''} w^{\varepsilon''} \to z'_w$ in L^2-weak as $\varepsilon'' \downarrow 0$. Following the above process once more, we obtain

$$(z'_w * \alpha)(x) = \int_Y z(x,y)w(y)\,dy \quad \text{a.e. in } \mathbb{R}^N. \tag{3.14}$$

By subtracting (3.14) from (3.13) we have

$$(z'_w - z_w) * \alpha = 0 \tag{3.15}$$

from which it follows that $z'_w = z_w$. Indeed the distributions (represented by the L^2 functions) α, $z'_w - z_w$ (respectively) have compact supports, i.e., they lie in $\mathscr{E}'(\mathbb{R}^N)$, the

subspace of $\mathscr{D}'(\mathbb{R}^N)$ formed of distributions having compact supports. But, since the vector space $\mathscr{E}'(\mathbb{R}^N)$ endowed with the convolution product is an algebra without zero divisor (see [8]), (3.15) implies $z'_w - z_w = 0$.

We have just established that for any subsequence ε' such that $u_{\varepsilon'} w^{\varepsilon'}$ converges weakly in L^2, the corresponding limit does not depend on ε'. That is, the sequence $u_\varepsilon w^\varepsilon$ converges weakly in L^2. The proof is complete. □

3.3. Extension of the first convergence result. Here and throughout the rest of § 3, ε denotes the subsequence involved in Proposition 2. Then, by that proposition, a unique function $z_w \in L^2$ is assigned to each w in C_p such that (3.3) holds. In other words, if we put

$$\Psi(x, y) = v(x)w(y) \quad \text{for } v \in \mathscr{K}(\mathbb{R}^N) \text{ and } w \in C_p \tag{3.16}$$

and $F_0(\Psi) = \int_{\mathbb{R}^N} z_w v \, dx$, we have $\int_{\mathbb{R}^N} u_\varepsilon \Psi^\varepsilon \, dx \to F_0(\Psi)$ for any Ψ in $\mathscr{K}_p$ of the form (3.16) (see (3.7) for the definition of Ψ^ε). This property is, clearly, what we call the first (or primitive) convergence result.

The aim in this section is then to extend the above property to all of $\mathscr{K}_p$.

LEMMA 3. *Let Ψ be fixed in $\mathscr{K}_p$ (Ψ independent of ε). Then the sequence $\varepsilon \to \int_{\mathbb{R}^N} u_\varepsilon \Psi^\varepsilon \, dx$ is Cauchy.*

Proof. Let $\Psi \in \mathscr{K}_p$. Let $\eta > 0$. Since the set $\mathscr{K}(\mathbb{R}^N) \otimes C_p$ is dense in $\mathscr{K}_p$ (see § 2), there exists some Ψ_η in $\mathscr{K}_p$, $\Psi_\eta = \sum_{i \in I} v_i \otimes w_i$ [$v_i \in \mathscr{K}(\mathbb{R}^N)$, $w_i \in C_p$], with I finite, such that the supports of both Ψ and Ψ_η lie in a fixed compact set $K \subset \mathbb{R}^N$ that depends only on Ψ, and

$$\sup_{x \in \mathbb{R}^N} \|\Psi_\eta(x) - \Psi(x)\|_{L^\infty} \leqq \frac{\eta}{2c} \tag{3.17}$$

where c is the constant in (3.1).

On the other hand, we evidently have for all ε

$$\sup_{x \in \mathbb{R}^N} |\Psi_\eta^\varepsilon(x) - \Psi^\varepsilon(x)| \leqq \sup_{x \in \mathbb{R}^N} \|\Psi_\eta(x) - \Psi(x)\|_{L^\infty}. \tag{3.18}$$

Now, consider ε_1, ε_2, destined to decrease independently. By a routine technique we have

$$\begin{aligned}
&\left| \int_{\mathbb{R}^N} u_{\varepsilon_2} \Psi^{\varepsilon_2} \, dx - \int_{\mathbb{R}^N} u_{\varepsilon_1} \Psi^{\varepsilon_1} \, dx \right| \\
&\quad \leqq \left| \int_{\mathbb{R}^N} u_{\varepsilon_2} (\Psi^{\varepsilon_2} - \Psi_\eta^{\varepsilon_2}) \, dx \right| + \left| \int_{\mathbb{R}^N} u_{\varepsilon_1} (\Psi_\eta^{\varepsilon_1} - \Psi^{\varepsilon_1}) \, dx \right| \\
&\qquad + \left| \int_{\mathbb{R}^N} u_{\varepsilon_1} \Psi_\eta^{\varepsilon_1} \, dx - \int_{\mathbb{R}^N} u_{\varepsilon_2} \Psi_\eta^{\varepsilon_2} \, dx \right|.
\end{aligned}$$

But (3.17) combines with (3.18) to give

$$\left| \int_{\mathbb{R}^N} u_{\varepsilon_i} (\Psi_\eta^{\varepsilon_i} - \Psi^{\varepsilon_i}) \, dx \right| \leqq \frac{\eta}{2} \quad \text{for } i = 1, 2.$$

Hence

$$\left| \int_{\mathbb{R}^N} u_{\varepsilon_2} \Psi^{\varepsilon_2} \, dx - \int_{\mathbb{R}^N} u_{\varepsilon_1} \Psi^{\varepsilon_1} \, dx \right| \leqq \eta + \left| \int_{\mathbb{R}^N} u_{\varepsilon_2} \Psi_\eta^{\varepsilon_2} \, dx - \int_{\mathbb{R}^N} u_{\varepsilon_1} \Psi_\eta^{\varepsilon_1} \, dx \right|.$$

Now, thanks to Proposition 2 we observe that for v in $\mathscr{K}(\mathbb{R}^N)$ and w in C_p the sequence $\varepsilon \to \int_{\mathbb{R}^N} u_\varepsilon v w^\varepsilon \, dx$ is Cauchy. Therefore, since $\int_{\mathbb{R}^N} u_\varepsilon \Psi_\eta^\varepsilon \, dx = \sum_{i \in I} \int_{\mathbb{R}^N} u_\varepsilon v_i w_i^\varepsilon$, the sequence $\varepsilon \to \int_{\mathbb{R}^N} u_\varepsilon \Psi_\eta^\varepsilon \, dx$ is Cauchy as a finite sum of Cauchy sequences. So we have $|\int_{\mathbb{R}^N} u_{\varepsilon_2} \Psi_\eta^{\varepsilon_2} \, dx - \int_{\mathbb{R}^N} u_{\varepsilon_1} \Psi_\eta^{\varepsilon_1} \, dx| \to 0$ as $\varepsilon_1 \downarrow 0$ and $\varepsilon_2 \downarrow 0$, and the conclusion follows from the arbitrariness of η. □

This brings us to one of the central preliminary convergence results in this work.

PROPOSITION 3. *For any $\Psi \in \mathscr{K}_p$ (Ψ independent of ε) there exists a unique real number $F_0(\Psi)$ such that*

$$\int_{\mathbb{R}^N} u_\varepsilon \Psi^\varepsilon \, dx \to F_0(\Psi) \quad \text{as } \varepsilon \downarrow 0.$$

3.4. End of the proof. Characterization of F_0. The aim in this section is to show that the above transformation $\Psi \to F_0(\Psi)$ is the restriction to $\mathscr{K}_p$ of a continuous linear form on $L^2(\mathbb{R}^N; L_p^2)$. More precisely, we must check that there exists a unique u_0 in $L^2(\mathbb{R}^N; L_p^2)$ such that

$$F_0(\Psi) = \int_{\mathbb{R}^N \times Y} u_0(x, y) \Psi(x, y) \, dx \, dy \quad \forall \Psi \text{ in } \mathscr{K}_p.$$

Since $\mathscr{K}_p$ is dense in $L^2(\mathbb{R}^N; L_p^2)$ and the transformation $\Psi \to F_0(\Psi)$ is linear, it suffices to establish that there exists a constant $c > 0$ such that

$$|F_0(\Psi)| \leqq c \|\Psi\|_{L^2(\mathbb{R}^N \times Y)} \quad \forall \Psi \text{ in } \mathscr{K}_p. \tag{3.19}$$

In this connection, fix Ψ in $\mathscr{K}_p$ (Ψ independent of ε). Then $|\int_{\mathbb{R}^N} u_\varepsilon \Psi^\varepsilon \, dx| \leqq c(\int_{K_0} |\Psi^\varepsilon|^2 \, dx)^{1/2}$ for all ε, where c is the constant on the right of (3.1).

By Proposition 3 and the fundamental property

$$\int_{K_0} |\Psi^\varepsilon|^2 \, dx \to \int_{K_0 \times Y} |\Psi(x, y)|^2 \, dx \, dy \quad \text{as } \varepsilon \downarrow 0 \quad \text{(see § 1)},$$

assertion (3.19) follows immediately. The proof is complete. □

Remark 4. The function u_0 has its support in the set $K_0 \times \mathbb{R}^N$ (or K_0, if u_0 is regarded as a function from $\mathbb{R}^N$ to L_p^2).

4. The leading-order approximation. A convergence theorem. In what follows, Ω denotes a bounded open set in the Euclidean space $\mathbb{R}^N$ (of the variables $x_1, \cdots, x_N$), Ω independent of ε. We denote by $\mathscr{K}(\bar{\Omega}; C_p)$ [respectively, $\mathscr{K}(\bar{\Omega})$] the set of all restrictions to Ω of functions in $\mathscr{K}_p$ [respectively, $\mathscr{K}(\mathbb{R}^N)$]. We also introduce the space $L^2(\Omega; L_p^2)$, which is a Hilbert space with the norm

$$\|u\|_{L^2(\Omega \times Y)} = \left[\int_{\Omega \times Y} |u(x, y)|^2 \, dx \, dy \right]^{1/2}.$$

The aim in this section is to establish the following theorem.

THEOREM 2. *Let $u_\varepsilon \in L^2(\Omega)$. Suppose that there exists a constant $c > 0$ such that*

$$\|u_\varepsilon\|_{L^2(\Omega)} \leqq c \quad \forall \varepsilon. \tag{4.1}$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, letting $\varepsilon \downarrow 0$,

$$\int_\Omega u_\varepsilon \Psi^\varepsilon \, dx \to \int_{\Omega \times Y} u_0(x, y) \Psi(x, y) \, dx \, dy \quad \forall \Psi \tag{4.2}$$

in $\mathcal{K}(\bar{\Omega}; C_p)$, *where* $u_0 \in L^2(\Omega; L^2_p)$. *Moreover,*

$$\int_\Omega u_\varepsilon v w^\varepsilon \, dx \to \int_{\Omega \times Y} u_0(x, y) v(x) w(y) \, dx \, dy \quad \forall v \tag{4.3}$$

in $\mathcal{K}(\bar{\Omega})$ *and all* w *in* L^2_p.

Proof. Property (4.2) is straightforward by Theorem 1 and Remark 4. As for (4.3), we begin by taking in (4.2) test functions of the form $\Psi(x, y) = v(x)w(y)$ with $v \in \mathcal{K}(\bar{\Omega})$, $w \in C_p$. We obtain as $\varepsilon \downarrow 0$,

$$\int_\Omega u_\varepsilon v w^\varepsilon \, dx \to \int_{\Omega \times Y} u_0(x, y) v(x) w(y) \, dx \, dy \tag{4.4}$$

for all $v \in \mathcal{K}(\bar{\Omega})$ and all $w \in C_p$.

Next, we must extend (4.4) to all functions w in L^2_p. Fix v in $\mathcal{K}(\bar{\Omega})$ and w in L^2_p. Let (w_n) be a sequence from C_p (dense subspace of L^2_p) such that $w_n \to w$ in L^2_p as $n \to \infty$. Utilizing the fact that the transformation $z \to z^\varepsilon$ is continuous linear from L^2_p to $L^2(\Omega)$ (see Remark 2), we have

$$\| w_n^\varepsilon - w^\varepsilon \|_{L^2(\Omega)} \leqq c_0 \| w_n - w \|_{L^2(Y)} \quad \forall n, \quad \forall \varepsilon < \varepsilon_0 \tag{4.5}$$

(c_0 and ε_0 are the constants in Lemma 2 with $K_0 = \bar{\Omega}$).

Now we write

$$\begin{aligned} &\int_\Omega u_\varepsilon v w^\varepsilon \, dx - \int_{\Omega \times Y} u_0 v w \, dx \, dy \\ &\quad = \int_\Omega u_\varepsilon v (w^\varepsilon - w_n^\varepsilon) \, dx + \int_{\Omega \times Y} u_0 v (w_n - w) \, dx \, dy \\ &\qquad + \int_\Omega u_\varepsilon v w_n^\varepsilon \, dx - \int_{\Omega \times Y} u_0 v w_n \, dx \, dy \end{aligned}$$

and estimate each of the first two integrals on the right-hand side separately (use (4.5)). This yields

$$\begin{aligned} &\left| \int_\Omega u_\varepsilon v w^\varepsilon \, dx - \int_{\Omega \times Y} u_0 v w \, dx \, dy \right| \\ &\qquad \leqq c_1 \| w_n - w \|_{L^2(Y)} + \left| \int_\Omega u_\varepsilon v w_n^\varepsilon \, dx - \int_{\Omega \times Y} u_0 v w_n \, dx \, dy \right| \end{aligned} \tag{4.6}$$

for all n and all $\varepsilon < \varepsilon_0$ (where c_1 is constant with respect to both ε and n).

Finally, let $\eta > 0$. Choose in (4.6) the natural number n so that $c_1 \| w_n - w \|_{L^2(Y)} \leqq \eta$. Then letting $\varepsilon \downarrow 0$ and using (4.4), it follows that the limit of the left-hand side of (4.6) is bounded from above by η. The desired conclusion then results from the arbitrariness of η.

Remark 5. Let u_ε be as in Theorem 2. First, let us observe that, by weak compactness, we may assume that in addition to (4.2) and (4.3) in Theorem 2, the subsequence ε satisfies the following property.

There exists $u \in L^2(\Omega)$ such that $u_\varepsilon \to u$ in $L^2(\Omega)$-weak. Next, taking $w = 1$ in (4.3) we easily obtain $u(x) = \int_Y u_0(x, y) \, dy$ (u is the mean value of u_0). It follows that u_0 is (uniquely) expressible in the form

$$u_0(x, y) = u(x) + \bar{u}_0(x, y) \quad \text{with} \int_Y \bar{u}_0(x, y) \, dy = 0.$$

So assume there is a subsequence from (u_ε) that converges strongly in $L^2(\Omega)$ as $\varepsilon \downarrow 0$. Then an easy computation yields $\bar{u}_0 = 0$; that is, the leading term u_0 in (1.2) does not depend on the local variables y. In other words, if the leading term depends on y, i.e., $\bar{u}_0 \neq 0$, then (u_ε) never contains a strongly convergent subsequence (see § 1).

5. The next-order approximation. A convergence theorem.

5.1. Notation and preliminaries. We denote by C_p^∞ the subspace of C_p formed of C^∞ functions, H_p^1 the subspace of L_p^2 formed of functions w such that $\partial w/\partial y_i \in L_p^2$ for $i = 1, \cdots, N$ (the derivatives obviously being taken in the distribution sense).

We provide H_p^1 with the norm

$$\|w\|_{H^1(Y)} = \left(\|w\|^2_{L^2(Y)} + \sum_{i=1}^N \left\| \frac{\partial w}{\partial y_i} \right\|^2_{L^2(Y)} \right)^{1/2},$$

which makes it a Hilbert space.

Sometimes it is more convenient to consider, instead of H_p^1, its closed subspace

$$\frac{H_p^1}{\mathbb{R}} = \left\{ w \in H_p^1; \int_Y w \, dy = 0 \right\}$$

on which the norm

$$\|w\|_{H^1(Y)/\mathbb{R}} = \left(\sum_{i=1}^N \left\| \frac{\partial w}{\partial y_i} \right\|^2_{L^2(Y)} \right)^{1/2}$$

is equivalent to the above H_p^1-norm.

We will need the following lemma.

LEMMA 4. *Let $f = (f_i), f_i \in L_p^2 (1 \leqq i \leqq N)$. Assume that $\sum_{i=1}^N \int_Y f_i w_i \, dy = 0$ for all $w = (w_i)$ in $(C_p^\infty)^N$ such that* div $w = 0$ *(where* div $w = \sum_{i=1}^N \partial w_i/\partial y_i$*). Then there exists a unique function $q \in H_p^1/\mathbb{R}$ such that $f_i = \partial q/\partial y_i$ for $i = 1, \cdots, N$.* □

Lemma 4 is the "periodic version" of the well-known result concerning the solvability of the equation grad $q = f$ for f given in $(L^2_{\text{loc}})^N$ (see, e.g., [10]). See, e.g., [6, Appendix] for the proof.

5.2. A convergence theorem (next-order approximation). We are now in a position to prove the main result in this section. In what follows, Ω denotes a smooth bounded open set in $\mathbb{R}^N$ (Ω independent of ε). As in the preceding sections, ε ($\varepsilon > 0$) denotes a sequence tending to zero.

THEOREM 3. *Let $u_\varepsilon \in H^1(\Omega)$. Suppose that there exists a constant $c > 0$ such that*

$$\|u_\varepsilon\|_{H^1(\Omega)} \leqq c \quad \forall \varepsilon. \tag{5.1}$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, as $\varepsilon \downarrow 0$,

$$u_\varepsilon \to u \text{ in } H^1(\Omega)\text{-weak}, \tag{5.2}$$

$$\int_\Omega \frac{\partial u_\varepsilon}{\partial x_i} \Psi^\varepsilon v \, dx \to \int_{\Omega \times Y} \left[\frac{\partial u}{\partial x_i}(x) + \frac{\partial u_1}{\partial y_i}(x, y) \right] \Psi(y) v(x) \, dx \, dy, \tag{5.3}$$

$i = 1, \cdots, N$; for all Ψ in L_p^2 and all v in $\mathscr{K}(\bar{\Omega})$, where $u_1 \in L^2(\Omega; H_p^1/\mathbb{R})$.

Proof. By virtue of (5.1) we can extract a subsequence such that (5.2) holds. Moreover, by Theorem 2 there exists $z_i \in L^2(\Omega; L_p^2)$, $1 \leqq i \leqq N$, such that

$$\int_\Omega \frac{\partial u_\varepsilon}{\partial x_i} \Psi^\varepsilon v \, dx \to \int_{\Omega \times Y} z_i(x, y) \Psi(y) v(x) \, dx \, dy \quad \text{as } \varepsilon \downarrow 0 \tag{5.4}$$

for all Ψ in L_p^2 and all v in $\mathscr{K}(\bar{\Omega})$.

It remains to show that there exists $u_1 \in L^2(\Omega; H^1_p/\mathbb{R})$ such that

$$z_i(x, y) = \frac{\partial u}{\partial x_i}(x) + \frac{\partial u_1}{\partial y_i}(x, y) \quad \text{for } i = 1, \cdots, N.$$

So let $\Psi = (\Psi_i)$ be a vector function in $(C^\infty_p)^N$ satisfying $\operatorname{div} \Psi = 0$. Then for v in $\mathcal{D}(\Omega)$ we have

$$\sum_{i=1}^N \int_\Omega \frac{\partial u_\varepsilon}{\partial x_i} \Psi_i^\varepsilon v \, dx = -\sum_{i=1}^N \int_\Omega u_\varepsilon \frac{\partial}{\partial x_i}(\Psi_i^\varepsilon v) \, dx.$$

By Leibniz's formula and the fact that $\operatorname{div} \Psi = 0$ (note that $\sum_{i=1}^N \partial \Psi_i^\varepsilon / \partial x_i = 1/\varepsilon (\operatorname{div} \Psi)^\varepsilon$) it follows that

$$\sum_{i=1}^N \int_\Omega \frac{\partial u_\varepsilon}{\partial x_i} \Psi_i^\varepsilon v \, dx = -\sum_{i=1}^N \int_\Omega u_\varepsilon \Psi_i^\varepsilon \frac{\partial v}{\partial x_i} \, dx.$$

By the Rellich theorem we may assume that the above subsequence ε satisfies the further property

$$u_\varepsilon \to u \quad \text{in } L^2(\Omega)\text{-strong},$$

so that, letting $\varepsilon \downarrow 0$ and recalling (5.4), we obtain

$$\sum_{i=1}^N \int_{\Omega \times Y} \left[z_i(x, y) - \frac{\partial u}{\partial x_i}(x) \right] \Psi_i(y) v(x) \, dx \, dy = 0$$

for all $\Psi \in (C^\infty_p)^N$, $\operatorname{div} \Psi = 0$, and all $v \in \mathcal{D}(\Omega)$. Hence we have for almost all $x \in \Omega$

$$\sum_{i=1}^N \int_Y \left[z_i(x, y) - \frac{\partial u}{\partial x_i}(x) \right] \Psi_i(y) \, dy = 0 \quad \text{for } \Psi \in (C^\infty_p)^N, \operatorname{div} \Psi = 0.$$

It follows by Lemma 4 that there exists a function u_1 from Ω to $H^1_p/\mathbb{R}$ such that

$$z_i(x, \cdot) - \frac{\partial u}{\partial x_i}(x) = \frac{\partial u_1(x)}{\partial y_i} \quad \text{a.e. in } \Omega (i = 1, \cdots, N). \tag{5.5}$$

Finally, from (5.5) we can easily show (e.g., by Lusin's characterization [3]) that u_1 is a measurable function from Ω to $H^1_p/\mathbb{R}$ (obtained from appropriate norm defined in § 5.1). Furthermore, again by (5.5) we have

$$\int_\Omega \|u_1(x)\|^2_{H^1(Y)/\mathbb{R}} \, dx < \infty$$

and the conclusion follows. □

6. A new approach in the theory of homogenization. Classically, the mathematical analysis of homogenization problems proceeds in two steps [1]. The first step, which is formal, derives, for example, from two-scale asymptotic expansions of the form

$$\begin{aligned} &u_\varepsilon(x) = u_0(x, y) + \varepsilon u_1(x, y) + \cdots, \quad y = \frac{x}{\varepsilon}, \\ &u_0, u_1, \cdots, \quad Y\text{-periodic in } y. \end{aligned} \tag{6.1}$$

More precisely, we postulate that the solution u_ε of a given problem (associated with a partial differential equation with coefficients ε-periodic) is similar to (6.1). Next, introducing (6.1) into the given problem yields a sequence of problems that determine $u_0, u_1, \cdots$.

The second step consists of rigorously proving the convergence of the preceding homogenization process, i.e., we must find some suitable topology in order that $\lim u_\varepsilon = u_0$ as $\varepsilon \to 0$. This validates the above formal calculations.

In this last section we propose an alternative approach. More precisely, we introduce a new asymptotic method for the mathematical analysis of homogenization problems. The method is quite straightforward. There is no need to postulate the existence of the functions u_0, u_1 in (6.1), since by Theorems 2 (or 1) and 3 such functions are available for a suitable subsequence from ε.

Our approach is illustrated by a regular homogenization problem. Nevertheless, the basic ideas can easily be extended to problems of the singular type.

6.1. Setting of the problem. In all that follows, unless otherwise specified, the summation convention is used.

Let Ω be a smooth bounded open set in $\mathbb{R}^N$ (the space of the variables $x_1, \cdots, x_N$) with boundary $\partial\Omega$. Let a_{ij} $(1 \leqq i, j \leqq N)$ be given functions defined on $\mathbb{R}^N$ (the space of the variables $y_1, \cdots, y_N$) and subject to the following conditions:

$$a_{ij} \in L^\infty, \quad a_{ij} \ Y\text{-periodic}, \quad a_{ij} = a_{ji}. \tag{6.2}$$

There exists $\alpha > 0$ such that the following holds for almost all y:

$$a_{ij}(y)\xi_i\xi_j \geqq \alpha|\xi|^2 \quad \forall \xi = (\xi_i) \in \mathbb{R}^N \tag{6.3}$$

(where the summation convention is utilized) with $|\xi|^2 = \sum_{i=1}^N \xi_i^2$.

Finally, let $f \in L^2(\Omega)$, and for each $\varepsilon > 0$ let u_ε be defined by

$$\begin{aligned} &u_\varepsilon \in H^1(\Omega), \\ &-\frac{\partial}{\partial x_i}\left(a_{ij}^\varepsilon \frac{\partial u_\varepsilon}{\partial x_j}\right) = f \quad \text{in } \Omega, \\ &u_\varepsilon = 0 \quad \text{on } \partial\Omega \end{aligned} \tag{6.4}$$

where $a_{ij}^\varepsilon(x) = a_{ij}(x/\varepsilon)$ (see (3.6)).

Clearly, from (6.2) (the first assumption) and (6.3), we see (6.4) uniquely determines u_ε.

Our aim is to find $\lim u_\varepsilon$ as $\varepsilon \downarrow 0$. In other words, we must study the homogenization problem associated with (6.4). Note that this problem has been solved in [1], where the results of the formal analysis were made rigorous by applying the Energy Method. As mentioned in § 1, we propose an alternative approach that should be more flexible and thus more adaptable for the study of unusual problems.

6.2. Description of the method. First, observe that u_ε, the solution of (6.4), satisfies

$$\begin{aligned} &u_\varepsilon \in H_0^1(\Omega), \\ &\int_\Omega a_{ij}^\varepsilon \frac{\partial u_\varepsilon}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx = \int_\Omega fv\,dx \quad \forall v \in H_0^1(\Omega). \end{aligned} \tag{6.5}$$

Next we estimate $\|u_\varepsilon\|_{H_0^1(\Omega)}$. Taking the particular test function $v = u_\varepsilon$ and using the boundedness and the coerciveness (from (6.3)) of the bilinear form in (6.5), we obtain

$$\|u_\varepsilon\|_{H^1(\Omega)} \leqq c \ (c > 0) \quad \forall \varepsilon.$$

Hence, the hypotheses of Theorem 3 are fulfilled. We can extract a subsequence still denoted by ε for simplicity such that

$$u_\varepsilon \to u \quad \text{in } H_0^1(\Omega)\text{-weak} \quad \text{as } \varepsilon \downarrow 0 \tag{6.6}$$

and, for all $\Psi \in L_p^2$, $v \in \mathscr{K}(\bar{\Omega})$,

$$\int_\Omega \frac{\partial u_\varepsilon}{\partial x_j} \Psi^\varepsilon v \, dx \to \int_{\Omega \times Y} \left[\frac{\partial u}{\partial x_j}(x) + \frac{\partial u_1}{\partial y_j}(x, y) \right] \Psi(y) v(x) \, dx \, dy, \quad j = 1, \cdots, N, \tag{6.7}$$

where $u_1 \in L^2(\Omega; H_p^1/\mathbb{R})$.

Derivation of the local problem. In (6.5) we take test functions of the form $v = \varepsilon w^\varepsilon \phi$ with $w \in H_p^1$, $\phi \in \mathscr{D}(\Omega)$. Then, noting that $\partial w^\varepsilon / \partial x_i = 1/\varepsilon (\partial w / \partial y_i)^\varepsilon$, where of course $(\partial w / \partial y_i)^\varepsilon(x) = \partial w / \partial y_i (x/\varepsilon)$, we are led to

$$\int_\Omega a_{ij}^\varepsilon \frac{\partial u_\varepsilon}{\partial x_j} \left(\frac{\partial w}{\partial y_i} \right)^\varepsilon \phi \, dx + \varepsilon \int_\Omega a_{ij}^\varepsilon \frac{\partial u_\varepsilon}{\partial x_j} w^\varepsilon \frac{\partial \phi}{\partial x_i} \, dx = \varepsilon \int_\Omega f w^\varepsilon \phi \, dx.$$

Now we propose passing to the limit as $\varepsilon \downarrow 0$. It is easy to check that both the second term on the left and the term on the right tend to zero. Hence,

$$\int_\Omega a_{ij}^\varepsilon \frac{\partial u_\varepsilon}{\partial x_j} \left(\frac{\partial w}{\partial y_i} \right)^\varepsilon \phi \, dx \to 0.$$

On the other hand, choose in (6.7) $\Psi = a_{ij}(\partial w / \partial y_i)$ (summation) with $w \in H_p^1$. By the above result we are finally led to

$$\int_{\Omega \times Y} a_{ij}(y) \left[\frac{\partial u}{\partial x_j}(x) + \frac{\partial u_1}{\partial y_j}(x, y) \right] \frac{\partial w}{\partial y_i}(y) \phi(x) \, dx \, dy = 0$$

for all $w \in H_p^1$ and all $\phi \in \mathscr{D}(\Omega)$. Hence the following holds for almost every x in Ω:

$$\int_Y a_{ij}(y) \left[\frac{\partial u}{\partial x_j}(x) + \frac{\partial u_1}{\partial y_j}(x, y) \right] \frac{\partial w}{\partial y_i}(y) \, dy = 0 \quad \forall w \in H_p^1. \tag{6.8}$$

Equation (6.8) is exactly that obtained by the formal method using multiple-scale asymptotic expansions (see [1]). It associates with the relation $u_1(x, \cdot) \in H_p^1/\mathbb{R}$ to give the so-called local problem, which permits us to express u_1 in terms of u. Evidently u_1 satisfies, for fixed x,

$$\begin{aligned} &u_1(x, \cdot) \in \frac{H_p^1}{\mathbb{R}}, \\ &\int_Y a_{ij} \frac{\partial u_1}{\partial y_j}(x, \cdot) \frac{\partial w}{\partial y_i} \, dy = -\frac{\partial u}{\partial x_j}(x) \int_Y a_{ij} \frac{\partial w}{\partial y_i} \, dy \quad \forall w \in \frac{H_p^1}{\mathbb{R}}, \end{aligned} \tag{6.9}$$

which is an elliptic variational problem for $u_1(x, \cdot)$, admitting one and only one solution.

We should stress that, contrary to the classical method, the resolution of the above problem does not concern us since u_1 has been constructed in Theorem 3. We only observe that u_1 is unique (i.e., independent of the subsequence extracted above) as soon as u is well determined.

Now we calculate u_1 in terms of u. Following [1], let χ^j $(j=1,\cdots,N)$ be defined by

$$\begin{aligned} &\chi^j \in \frac{H^1_p}{\mathbb{R}}, \\ &\int_Y a_{kh}\frac{\partial \chi^j}{\partial y_h}\frac{\partial w}{\partial y_k}\,dy = \int_Y a_{kj}\frac{\partial w}{\partial y_k}\,dy \quad \forall w \in \frac{H^1_p}{\mathbb{R}}. \end{aligned} \tag{6.10}$$

Then, from the preceding remarks, we see that u_1 is given by

$$u_1(x,y) = -\frac{\partial u}{\partial x_j}(x)\chi^j(y). \tag{6.11}$$

Indeed, the function on the right-hand side is the solution of (6.9).

Remark 6. It is not difficult to verify that the equation in (6.10) can be written under the form

$$a(\chi^j - y_j, w) = 0 \quad \forall w \in H^1_p/\mathbb{R} \tag{6.12}$$

where $a(\cdot,\cdot)$ is the bilinear form that figures in (6.10), and $y_j (1 \leqq j \leqq N)$ are the coordinate functions. □

Derivation of the global (or limit) problem. The point now is to derive the boundary value problem satisfied by the global limit u. This is straightforward. Choose in (6.7) the particular function $\Psi = a_{ij}$, and in place of the v's consider the derivatives $\partial v/\partial x_i$, with $v \in \mathcal{D}(\Omega)$. Hence, summing over i,j on both sides of (6.7) and using (6.5) yields

$$\int_{\Omega\times Y} a_{ij}(y)\left[\frac{\partial u}{\partial x_j}(x) + \frac{\partial u_1}{\partial y_j}(x,y)\right]\frac{\partial v}{\partial x_i}(x)\,dx\,dy = \int_\Omega fv\,dx,$$

which by (6.11) becomes

$$\int_\Omega \left(\tilde{a}_{ij} - \int_Y a_{ih}\frac{\partial \chi^j}{\partial y_h}\,dy\right)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx = \int_\Omega fv\,dx$$

where $\tilde{a}_{ij} = \int_Y a_{ij}(y)\,dy$. But it is easy to check that

$$\tilde{a}_{ij} - \int_Y a_{ih}\frac{\partial \chi^j}{\partial y_h}\,dy = -a(y_i, \chi^j - y_j).$$

On the other hand, by (6.12) we have easily that $-a(y_i, \chi^j - y_j) = a(\chi^i - y_i, \chi^j - y_j)$ (note that the form $a(\cdot,\cdot)$ is symmetric). From all that we deduce the problem for u:

$$\begin{aligned} &u \in H^1_0(\Omega), \\ &\int_\Omega q_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx = \int_\Omega fv\,dx \quad \forall v \in H^1_0(\Omega), \end{aligned} \tag{6.13}$$

where

$$q_{ij} = a(\chi^i - y_i, \chi^j - y_j). \tag{6.14}$$

The constants q_{ij} are the so-called *homogenized coefficients.* They satisfy the ellipticity condition

$$q_{ij}\xi_i\xi_j \geqq c|\xi|^2 (c>0) \quad \forall \xi \in \mathbb{R}^N \quad \text{(see [1])}$$

so that u is uniquely determined by (6.13). Consequently, the subsequence ε in (6.6) and (6.7) may be replaced by the whole sequence from which it was extracted.

Thus, we have proved the following homogenization theorem.

THEOREM 4. *For each $\varepsilon > 0$ let u_ε be the solution of the boundary value problem* (6.4). *Then, as $\varepsilon \downarrow 0$,*

$$
\begin{aligned}
&u_\varepsilon \to u \quad \text{in } H_0^1(\Omega)\text{-weak},\\
&\int_\Omega \frac{\partial u_\varepsilon}{\partial x_j} \Psi^\varepsilon v \, dx \to \left(\int_\Omega \frac{\partial u}{\partial x_k} v \, dx \right) \int_Y \Psi \frac{\partial}{\partial y_j} (y_k - \chi^k) \, dy,
\end{aligned} \tag{6.15}
$$

$j = 1, \cdots, N$, for all $\Psi \in L_p^2$ and all $v \in \mathscr{K}(\bar{\Omega})$, where u is the solution of the boundary value problem

$$
\begin{aligned}
-\frac{\partial}{\partial x_i}\left(q_{ij} \frac{\partial u}{\partial x_j} \right) &= f \quad \text{in } \Omega\\
u &= 0 \quad \text{on } \partial\Omega,
\end{aligned}
$$

q_{ij} given by (6.14), *and $\chi^k (k = 1, \cdots, N)$ given by* (6.10).

Remark 7. The usual homogenization theorem [1], [9] does not involve (6.15). This property obviously follows from (6.7) and (6.11).

6.3. Concluding remarks. We have just proposed a new asymptotic method for the mathematical analysis of homogenization problems. The method is straightforward and quite natural. Note that, although we have chosen a problem of the regular type for a detailed analysis, our approach is essentially based on Theorem 1, which requires only the weaker assumption that u_ε remains bounded in L^2. So it is reasonable to assume the above idea can be successfully extended to a more general situation involving problems of the singular type.

Acknowledgment. The author thanks the referees for their kind advice.

REFERENCES

[1] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[2] N. BOURBAKI, *Eléments de Mathématiques*, Vol. XIII, *Intégration*, Hermann, Paris, 1965.

[3] R. E. EDWARDS, *Functional Analysis*, Holt, Rinehart and Winston, New York, 1965.

[4] F. FLEURY, *Propagation des ondes dans une suspension de particules solides*, Compt. Rend. Acad. Sci. Paris, sér. A, 288 (1979), pp. 77–80.

[5] T. LEVY, *Propagation of waves in a fluid-saturated porous elastic solid*, Internat. J. Engrg. Sci., 17 (1979), pp. 1005–1014.

[6] G. NGUETSENG, *Sur quelques problèmes de perturbations dans des ouverts périodiques et applications à la mécanique des composites*, Thèse d'Etat, Université Paris 6, 1984.

[7] E. SANCHEZ-PALENCIA, *Non-Homogeneous Media and Vibration Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1980.

[8] L. SCHWARTZ, *Théories des distributions*, Hermann, Paris, 1966.

[9] L. TARTAR, *Problèmes d'homogenization dans les équations aux dérivées partielles*, Cours Peccot (rédigé par F. Murat), Collège de France, 1977.

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

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 624-633, May 1989

009

A NONABELIAN VERSION OF THE SHANNON SAMPLING THEOREM*

A. H. DOOLEY†

Abstract. Using techniques from the theory of contraction of Lie groups, a version of the Shannon sampling theorem appropriate to Cartan motion groups is proved. This work has possible applications in digital signal analysis.

Key words. contractions, sampling theorems, Hankel transform

AMS(MOS) subject classifications. 22E46, 42C99

1. Introduction. The Shannon sampling theorem for the real line R may be stated as follows.

THEOREM A. *Let $f \in L^2(R)$ have Fourier transform $\hat{f}(\nu) = 1/2\pi \int f(x)\, e^{-i\nu x}\, dx$, and suppose that $\hat{f}$ is supported in $[-S, S]$. Then*

$$f(x) = \sum_{n=-\infty}^{\infty} d_n(x) f\left(\frac{n}{2S}\right),$$

where $d_n(x) = k_S(x-(n/2S)) = 1/2S \int_{-S}^{S} e^{i(x-(n/2S))\nu}\, d\nu$.

Thus, knowledge of f at the points $n/2S$, $n = 0, \pm 1, \cdots$, allows reconstruction of f everywhere, by use of the kernels k_S.

The obvious generalization of this theorem to $L^2(R^2)$ holds, where the two-dimensional Fourier transform is considered. However, we do not wish to pursue this line here. Rather, we want to discuss a version of the sampling theorem appropriate to the Fourier-Bessel transform of $f \in L^2(R^2)$, which is given by considering f as a linear combination of $F_n(re^{i\alpha}) = \phi_n(r)\, e^{in\alpha}$. The Fourier-Bessel transform of F_n is given by

$$\hat{F}_n(\mathrm{Re}^{i\psi}) = 2\pi i^n\, e^{in\psi} \Phi_n(R),$$

where

$$\Phi_n(R) = \int_0^{\infty} \phi_n(r) J_n(Rr) r\, dr.$$

(Here, J_n denotes the nth Bessel function.)

The following theorem is due to Kramer [9].

THEOREM B. *Suppose $f \in L^2(R^2)$. If $\int_0^{2\pi} f(re^{i\phi})\, e^{ik\phi}\, d\phi = 0$ for $|k| \geqq M$ and if there is $R_0 > 0$ such that for all k, $\Phi_k(R) = 0$ whenever $R > R_0$, then*

$$f(re^{i\phi}) = \sum_{k=-M}^{M} e^{ik\phi} \sum_{n=1}^{\infty} f_k(t_{k,n}) S_{k,n}(r)$$

where $S_{k,n}(r) = 2t_{k,n} J_k(r)/((t_{k,n}^2 - r^2) J_{k+1}(t_{k,n}))$. In these expressions, $(t_{k,n})_{n=1}^{\infty}$ are the zeros of the Bessel function J_k.

Kramer's theorem applies not only to expansions in Bessel functions, but also to a number of other special function expansions—see Jerri [6, § 3] for a full discussion of these generalizations. Analogues of Theorem B may be obtained for these expansions, but the above version will suffice to explain the background to the present paper.

* Received by the editors February 24, 1987; accepted for publication (in revised form) July 5, 1988. This research was supported by the Australian Research Grants Scheme.

† School of Mathematics, University of New South Wales, Kensington, New South Wales, 2033, Australia.

Theorem B is somewhat unsatisfactory in certain respects. First, from a practical point of view, the data are rarely presented at the zeros of the Bessel functions—these are hard to locate, too. Secondly, the sampling must be at zeros of different Bessel functions for each component f_k of f. Thirdly, the comparatively neat form of the kernel $S_n(r)$ is obtained by using rather particular properties of J_k not shared for other special functions; the kernels obtained by Jerri [7] are much more complicated.

I would like to propose here an alternative theorem based on the Fourier-Bessel transform in R^2 (and more generally on the Radon transform in R^n) but in which the sample points are spaced equally. Of course, we can no longer expect to match f exactly—it is known that the classical Shannon sampling is the only way to do that. However, the L^2 difference between the reconstructed function and f can be estimated uniformly in the spacing.

Our general theorem applies to the Fourier transform on a motion group (Theorem 2.1). A corollary (Theorem 4.2) deals with the Radon transform on the tangent space of a symmetric space. For the purposes of comparison, we present here the simple case of R^2 considered relative to the rotation action of $S0(2)$. This is a direct analogue of Theorem B where the sampling is done at equally spaced points (grid $1/\lambda$) and f is reconstructed to within $0(1/\lambda)$.

THEOREM C. *Let f satisfy the conditions of Theorem B. Then*

$$f(re^{i\psi}) = \sum_{k=-m}^{m} e^{ik\psi} \sum_{n=1}^{\infty} (2n+1) h_\lambda^k\left(r, \frac{n}{\lambda}\right) \Phi_k\left(\frac{n}{\lambda}\right) + 0\left(\frac{1}{\lambda}\right)$$

where

$$h_\lambda^k\left(r, \frac{n}{\lambda}\right) = \frac{1}{\lambda^2} \int_0^\lambda J(Sr) J\left(S\frac{n}{\lambda}\right) S\, dS, \text{ and } \left\| 0\left(\frac{1}{\lambda}\right) \right\|_2 \leqq \frac{K}{\lambda}.$$

The final section of this paper details some applications of these abstract theorems to the theory of signal processing.

2. Preliminaries. Let (G, K) be a Riemannian symmetric pair of the compact type. (We will recover Theorem C in the case $G = S0(3)$, $K = S0(2)$.) Then [4] we may write $\mathbf{g} = \mathbf{k} + V$, where V is a vector space complement for $\mathbf{k}$ in $\mathbf{g}$, and we may form the Cartan motion group associated with (G, K). This is the semidirect product $V \rtimes K$ of V by the adjoint action of K. Of particular interest will be sampling for functions on $V \rtimes K$ which are bi-K-invariant—these are in bijective correspondence with functions on V which are K-invariant. In the special case of $(G, K) = (S0(n+1), S0(n))$, this will amount to an approximate sampling theorem for rotationally invariant functions on R^n, i.e., for the n-Hankel transform on R^+.

Central to our analysis will be techniques from [2], where an approximation theorem for matrix coefficients was obtained [2, Thm. 2]. The map $\pi_\lambda : V \rtimes K \to G$ defined by $\pi_\lambda(v, k) = \exp_G{}^{v/\lambda} \cdot k$ generalizes the map $R \to T : r \to e^{ir/\lambda}$, and gives the structural link between the groups which enables us to "transfer" harmonic analysis.

I can now explain the idea behind the proof of Theorem 1. Consider the following proof of Theorem A, which has been attributed to Kolmogorov.

Let $f \in L^1 \cap L^2(R)$ and suppose that $\operatorname{supp} \hat{f} \subseteq [-S, S]$. Then $f(x) = \int_{-S}^{S} \hat{f}(w)\, e^{iwx}\, dw$. For each x we may expand $w \to e^{iwx}$ in a Fourier series on $[-S, S]$, obtaining

$$e^{iwx} = \sum_{n=-\infty}^{\infty} d_n(x)\, e^{i(n/2S)w},$$

where $d_n(x)$ is as in Theorem A of the Introduction. Substituting back in, we have

$$f(x)=\sum_{n=-\infty}^{\infty} d_n(x)\int \hat{f}(w)\,e^{i(n/2S)w}\,dw=\sum_{n=-\infty}^{\infty} d_n(x)f\left(\frac{n}{2S}\right).$$

The idea is to generalize the above proof, replacing R by $V\rtimes K$ and $[-S, S]$ by the group G, via the contraction map π_λ.

A set of irreducible representations of $V\rtimes K$ of full Plancherel measure may be described as follows (cf. [1], [8]). Choose a maximal Abelian subalgebra $\mathbf{a}$ of V, and let M be the stabilizer of $\mathbf{a}$ in K. Given $\psi\in\mathbf{a}^{*+}$ and $\mu\in\hat{M}$ acting in $\underset{\sim}{H}_\mu$ we have a representation $\rho_{\mu,\psi}$ acting in

$$\underset{\sim}{K}_\mu=\{h\in L^2(K,\underset{\sim}{H}_\mu)\colon \mu(m)h(km)=h(k),\ \forall\, m\in M,\ \forall k\in K\}$$

by $(\rho_{\mu,\psi}(v,k)h)(k_0)=e^{i\psi(Ad(k_0^{-1})v)}h(k^{-1}k_0)$.

For $f\in L^1(V\rtimes K)$, we define $\hat{f}(\rho_{\mu,\psi})\in B(\underset{\sim}{H}_{\mu,\psi})$ by

$$(\hat{f}(\rho_{\mu,\psi})h)(k_0)=\int_V\int_K e^{i\psi(Adk_0^{-1}V)}\,h(k^{-1}k_0)f(v,k)\,du\,dk.$$

The Plancherel formula for $V\rtimes K$ (cf. [7]) then states that for $f\in C_c^\infty(V\rtimes K)$

$$f(x)=\sum_{\mu\in\hat{M}} d_\mu\int_{(\mathbf{a}^*)^+}\operatorname{Tr}\left(\hat{f}(\rho_{\mu,\phi})\rho_{\mu,\phi}(x)\right)\prod_{\alpha\in P_+}\phi(H_\alpha)\,d\phi,$$

where P_+ denotes a set of reduced roots for G on $\mathbf{a}$.

Given $\mu\in\hat{M}$, we set

$$f_\mu(x)=\int_{(\mathbf{a}^*)^+}\operatorname{Tr}\left(\hat{f}(\rho_{\mu,\phi})\rho_{\mu,\phi}(x)\right)\prod_{\alpha\in P_+}\phi(M_\alpha)\,d\phi.$$

(The f_μ are the analogues of the F_n in § 1.)

Let P denote the set of K-class one weights of $\mathbf{a}$ in $\mathbf{a}^*$. We may now state our version of the sampling theorem.

THEOREM 2.1. *Let $f\in L^1(V\rtimes K)$ and suppose that*

(i) $\operatorname{supp}\hat{f}\subseteq\{\rho_{\mu,\phi}\colon |\phi|\leqq R\}$

(ii) *$\hat{f}(\rho_{\mu,\phi})$ has finite rank as an operator on $\underset{\sim}{H}_{\mu,\phi}$.*

Then $f=\sum_{\mu\in\hat{M}} d_\mu f_\mu$, and each f_μ may be approximately reconstructed by, for $v=k_0\eta\in V$,

$$f_\mu(v,k)=\sum_{\beta\in P} d_{\mu,\beta}\sum_{r,s,t,\ell=0}^{M} K^{r,s,t,\ell}\left(\eta,\frac{\beta}{\lambda}\right)f_{s,r,\ell,t}\left(k_0\frac{\beta}{\lambda},k\right)+0\left(\frac{1}{\lambda}\right).$$

Here, for a suitable choice of basis for $\underset{\sim}{K}_\mu$ (specified below),

$$K^{r,s,t,\ell}_{\mu,\lambda}\left(\phi,\frac{\beta}{\lambda}\right)=\lambda^{\dim V}\int_{B_\lambda}(\rho_{\mu,\phi}(x))_{r,s}\,\overline{(\rho_{\mu,\,i\beta/\lambda}(x))_{t,\ell}}\,dx$$

and

$$f_{s,r,\ell,t}(x)=\int_{|\phi|\leqq R}\hat{f}(\rho_{\mu,\phi})_{s,r}(\rho_{\mu,\phi}(x))_{\ell,t}\prod_{\alpha\in P_+}\phi(H_\alpha)\,d\phi.$$

The $0(1/\lambda)$ is uniform on compact subsets of $V\rtimes K$.

This theorem will be proved in the next section.

3. Proof of the theorem. Before giving the proof, we will need some lemmas. Let $f \in L^1 \cap L^2(V \rtimes K)$ and suppose $\operatorname{supp} \hat{f} \subseteq \{\rho_{\mu,\phi}: |\phi| \leqq R\} = S_{\mu,R}$. We then have

$$f_\mu(x) = \int_{\{\phi: |\phi| \leqq R\}} \operatorname{Tr}(\hat{f}(\rho_{\mu,\phi})\rho_{\mu,\phi}(x)) \prod_{\alpha \in P_+} \phi(H_\alpha)\, d\phi.$$

We may write $x \in V \rtimes K$ as $x = (v, k)(v \in V, k \in K)$ and using the fact ([3]) that each K-orbit in V intersects $\mathbf{a}^+$ in just one point, we may write $v = k_0\beta$, $\beta \in \mathbf{a}^+$, and $k_0 \in K$. Our first lemma enables us to "pass β from the group to its dual." We identify $\mathbf{a}$ and $\mathbf{a}^*$ by the Killing form $\langle , \rangle$.

LEMMA 3.1. *Let $\phi \in \mathbf{a}$, $\beta \in \mathbf{a}^*$, $\mu \in \hat{M}$, $k, k_0 \in K$. Then*

$$\rho_{\mu,\phi}(k_0\beta, k) = \rho_{\mu,\beta}(k_0\phi, k).$$

Proof. This is simply a matter of looking at the definition of $\rho_{\mu,\phi}$. □

Using this lemma, we may write

$$f_\mu(k_0\beta, k) = \int_{|\phi| \leqq R} \operatorname{Tr}(\hat{f}(\rho_{\mu,\phi})\rho_{\mu,\beta}(k_0\phi, k)) \prod_{\alpha \in \Phi_+} \phi(H_\alpha)\, d\phi. \tag{1}$$

In order to analyse the trace, we choose an orthonormal basis for $\underset{\sim}{K}_\mu$ as follows (cf. [1, § 4]). Take a sequence $\beta_1, \beta_2 \cdots$ of vectors in the lattice P of K-class one vectors in $\mathbf{a}^{*+}$ such that

$$\left|\beta - \frac{1}{n}\sum_{j=1}^{n} \beta_j\right| = 0\left(\frac{1}{n}\right).$$

The vectors $\mu + \psi_n = \mu + \sum_{j=1}^{n} \beta_j \in \mathbf{t}^*$ define integral weights and hence representations σ_{μ,ψ_n} of G acting in

$$\underset{\sim}{H}_{\mu,\psi_n} = \{f \in C^\infty(G): \text{(i) } f(gt) = e^{i(\psi_n, t)} f(g); \text{(ii) } Xf = 0\ \forall X \in \eta_+\}.$$

(Here $\eta_+ = \bigoplus_{\alpha \in P_+} \mathbf{g}_\alpha$.)

The images $\underset{\sim}{H}_{\mu,\psi_n}|_K$ form an ascending union of subspaces that are dense in $\underset{\sim}{K}_\mu$; we denote by $R = R_{\mu,n}$ the restriction map $f \to f|_K$. We choose an orthonormal basis $\{u_i\}$ for K_μ, which is compatible with this structure in the sense that there exists a sequence $i(1), i(2) \cdots$ such that $u_{i(j)}, u_{i(j)}, \cdots, u_{i(j+1)-1}$ is an orthonormal basis for $\underset{\sim}{H}_{\mu,\psi_0}|_K$.

For $A \in B(\underset{\sim}{H}_\mu)$, we will denote by $A_{r,s}$ the components of A with respect to this basis, i.e.,

$$A_{r,s} = (Au_r, u_s)_{L^2(K)}.$$

The assumption that $\hat{f}(\rho_{\mu,\phi})$ has finite rank implies that there exists $M \in N$ such that for all ϕ,

$$\hat{f}(\rho_{\mu,\phi})_{r,s} = 0 \quad \forall\ r, s \geqq M.$$

Thus

$$\operatorname{Tr}(\hat{f}(\rho_{\mu,\phi})\rho_{\mu,i\beta}(k_0\phi, k)) = \sum_{r,s=1}^{M} (\rho_{\mu,\phi})_{r,s}(\rho_{\mu,i\beta}(k_0\phi, k))_{s,r}. \tag{2}$$

In order to demonstrate our estimate, we need an approximate identity on $V \rtimes K$.

LEMMA 3.2. *For each $\delta > 0$, there is a nonnegative function $\kappa_\delta \in C^\infty(G)$ with the following properties*

i) $\operatorname{supp} \kappa_\delta \subseteq \pi_1(V \rtimes K)$

ii) $\hat{\kappa}_\delta$ *is rapidly decreasing on* G

iii) $\kappa_{\delta,\lambda} = \lambda^{\dim V}(\kappa_\delta \otimes \pi_\lambda)$ *is an approximate identity for* $L^2(V \rtimes K)$ *as* $\delta \to 0$ *and* $\lambda \to \infty$

iv) *For all* μ *and for all* $\varepsilon > 0$, *there exists* $\delta_0 > 0$ *such that whenever* $\delta < \delta_0$, $|\kappa^*_{\delta,\lambda}(\rho_{\mu,\eta})_{r,s}(v,k) - (\rho_{\mu,\eta})_{r,s}(v,k)| < 0(1/\lambda) + \varepsilon$ *(as* $\lambda \to \infty$*) for all* $r, s = 1, \cdots, M$ *for all* η *and for all* v.

Proof. Let $B \subseteq V$ be an $Ad(K)$-invariant set such that $\exp_G$ is one-to-one on B. Choose a nonnegative (and nonzero) C^∞ function $\tilde{\tilde{\kappa}}$ on V which is $Ad(K)$-invariant and supported inside B. Let $(h_\delta)_{\delta \to 0}$ be a nonnegative approximate identity for K (see [5], 25.47).

Set $\tilde{\kappa} = \tilde{\tilde{\kappa}} \circ \log_G$ and let κ_δ be defined by $\kappa_\delta(g) = (\tilde{\kappa} \otimes h_g)(\pi_1^{-1}(g))$. (Note that π_1 is one-to-one on $B \times K$, cf. [2, § 3].) Here, I denote by $\tilde{\kappa} \otimes h_\delta$ the function $(v,k) \to \tilde{\kappa}(v)h_\delta(k)$.

Since $\kappa_\delta \in C_c^\infty$, its Fourier transform is rapidly decreasing. An easy calculation shows that

$$\lambda^{\dim V}(\kappa_\delta \circ \pi_\lambda)(v,k) = \lambda^{\dim V}\tilde{\kappa}\left(\frac{v}{\lambda}\right)h_\delta(k).$$

The latter is clearly an approximate identity for $L^2(V \rtimes K)$; in fact, letting $\tilde{\kappa}_\lambda(v) = \lambda^{\dim V}\tilde{\kappa}(v/\lambda)$, we have

$$((\tilde{\kappa}_\lambda \otimes h_\delta) * f)(v,k) = \int\int \tilde{\kappa}_\lambda(v - v_1)h_\delta(k_1^{-1}k)f(v_1, k_1)\, dv_1\, dk_1.$$

To prove (iv), we calculate in the same manner that

$$(\kappa_{\delta,\lambda} * \rho_{\mu,\eta})_{r,s}(v,k) = \langle \tilde{\kappa}_\lambda * e^{i\eta k_0^{-1}(\cdot)}(v)h_\delta * u_r(k^{-1}k_0), u_s(k_0)\rangle H_\mu.$$

By a suitable choice of δ, we have $h_\delta * u_r$ arbitrarily close to u_r in $\underset{\sim}{K}_\mu$. On the other hand, $\tilde{\kappa}_\lambda * e^{i\eta k_0^{-1}(\cdot)}(v) = \tilde{\kappa}_\lambda * e^{i\eta(\cdot)}(k_0 v)$, and we calculate that $\tilde{\kappa}_\lambda * e^{i\eta(\cdot)}(v) = e^{i\eta v}\int_V \tilde{\kappa}(v_1)\, e^{-i\eta v_1 \lambda}\, dv_1$.

Since we are integrating over a bounded set B, we see that $|\kappa_\lambda * e^{i\eta(\cdot)}(v) - e^{i\eta v}| \leqq 0(1/\lambda)$. This gives (iv). □

LEMMA 3.3. *For all* $g \in G$, $\mu \in \hat{M}$, $\phi \in \mathfrak{a}^{*+}$ *and* $1 \leqq s, r \leqq M$,

$$\kappa_\delta * (\rho_{\mu,\phi} \otimes \pi_\lambda^{-1})_{s,r}(g) = \sum_{\sigma \in \hat{G}} d_\sigma \operatorname{Tr}(\hat{\kappa}_\delta(\sigma)(\rho_{\mu,\phi} \otimes \pi_\lambda^{-1})_{s,r})\hat{}(\sigma)\sigma(g).$$

Proof. This holds since $\kappa_\delta \in C^\infty(G)$. □

We next summarize, in a form convenient to the present article, some results from [1] and [2].

LEMMA 3.4. *Let* $\sigma \in \hat{G}$, $\sigma = \sigma_{\alpha+\beta}$ *where* $\alpha \in P_-$ *and* $\beta \in P_+$. *Then for all* $\phi \in \mathfrak{a}^*$

$$(\rho_{\mu,i\phi} \otimes \pi_\lambda^{-1})^\wedge_{r,s}(\sigma) = 0 \quad \textit{unless } \alpha = \mu.$$

Proof. Let T_1 be a maximal torus in M. Then $(\rho_{\mu,i\phi} \otimes \pi_\lambda^{-1})_{r,s}$ transforms as χ_μ under the left action of T_1. On the other hand, $\sigma_{p,q}(g)$ transforms as χ_α under left action of T_1. Thus unless $\alpha = \mu$ we have orthogonality. □

LEMMA 3.5. *There is a constant* $M = M(A, R)$ *such that whenever* $|\beta/\lambda - \phi| < A/\lambda (\beta \in P, \phi \in \mathfrak{a}^{*+})$ *and* $(v,k) \in B_R$,

$$\|R\sigma_{\mu,\beta}(\pi_\lambda(v,k))R^{-1}u - \rho_{\mu,\phi}(v,k)u\|_{H_\mu} \leqq \frac{M}{\lambda}.$$

Proof. This follows immediately from [2, Thm. 3]. □

Hence

$$|\phi_{\mu,\beta}(\pi_\lambda(v,k))_{t,\ell} - \rho_{\mu,\phi}(v,k)_{t,\ell}| < \frac{M}{\lambda}. \tag{3}$$

If we take into account the definition of $K_{\mu,\lambda}^{r,s,t,\ell}$ (see Theorem 2.1), Lemma 3.6 follows immediately from (3).

LEMMA 3.6. *For* $1 \leqq \ell, s, r, t \leqq M$,

$$\left| ((\rho_{\mu,\phi} \circ \pi_\lambda^{-1})^\wedge_{s,r}(\sigma_{\mu,\beta}))_{t,\ell} - K_\mu^{r,s,t,\ell}\left(\sigma, \frac{\beta}{\lambda}\right) \right| < 0\left(\frac{1}{\lambda}\right)$$

where "0" *depends on R and M.*

Now

$$(\rho_{\mu,\phi} \circ \pi_\lambda^{-1})^\wedge_{s,r}(\sigma_{\mu,\beta})_{t,\ell} = \int_G (\rho_{\mu,\phi}(\pi_\lambda^{-1}(g)))_{s,r} \overline{(\sigma_{\mu,\beta}(g))_{t,\ell}}\, dg$$

and by [1] Lemma 3.3, this is equal to

$$\frac{1}{\lambda^{\dim V}} \int_{B_\lambda} (\rho_{\mu,\phi}(x))_{s,r} (\sigma_{\mu,\beta}(\pi_\lambda(x))_{t,\ell} dx + 0\left(\frac{1}{\lambda^{2+\dim V}}\right)$$

where $B_\lambda = \{(v,k) \in V \rtimes K : |v| \leqq \lambda\}$.

Now according to (3), the latter expression is, to within $0(1/\lambda)$, equal to

$$\frac{1}{\lambda^{\dim V}} \int_{B_\lambda} (\rho_{\mu,\phi}(x))_{s,r} (\rho_{\mu,\beta/\lambda}(x))_{t,\ell}\, dx$$

which is by definition $K_{\mu,\lambda}^{s,r,t,\ell}(\phi, B/\lambda)$. □

We may now prove Theorem 2.1.

Proof of the theorem. By (1) and (2), we have

$$f_\mu(x) = \sum_{r,s=1}^{M} \int_{|\phi| \leqq R} \hat{f}(\rho_{\mu,\phi})_{r,s} [\rho_{\mu,\eta}(k_0\phi, k)]_{s,r} \prod_{\alpha \in P_+} \phi(H_\alpha)\, d\phi. \tag{3}$$

Using Lemma 3.2 (iv), we have

$$|\lambda^{\dim V}(\kappa_\delta \circ \pi_\lambda) * (\rho_{\mu,\eta})_{r,s}(k_0\phi, k) - (\rho_{\mu,\eta})_{r,s}(k_0\phi, k)| < 0\left(\frac{1}{\lambda}\right) + \varepsilon \text{ as } \lambda \to \infty.$$

Furthermore, by Lemma 3.3 of [2]

$$|\lambda^{\dim V}((\kappa_\delta \circ \pi_\lambda) * (\rho_{\mu,\eta})_{r,s})(k_0\phi, k) - (\kappa_\delta * [(\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1}])(\pi_\lambda(k_0\phi, k))| < 0\left(\frac{1}{\lambda}\right). \tag{4}$$

(Note that the first convolution in (3) is in $V \rtimes K$ whereas the second is in G.) Equation (4) holds uniformly for $|\phi| < R$.

From (3) and (4) we deduce

$$|\rho_{\mu,\eta}(k_0\phi, k)_{r,s} - (\kappa_\delta * (\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1})(\pi_\lambda(k_0\phi, k))| < 0\left(\frac{1}{\lambda}\right). \tag{5}$$

Now Lemmas 3.3 and 3.4 give

$$\begin{aligned}
&(\kappa_\delta * (\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1})(\pi_\lambda(k_0\phi, k)) \\
&\quad = \sum_{\sigma \in \hat{G}} d_\sigma \operatorname{Tr}\{\hat{\kappa}_\delta(\sigma)((\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1})^\wedge(\sigma)\sigma(\pi_\lambda(k_0\phi, k))\} \\
&\quad = \sum_{\beta \in P} d_{\mu,\beta} \sum_{t,\ell=0}^{\min(M, d_{\mu,\beta})} (\hat{\kappa}_\delta(\sigma_{\mu,\beta})((\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1})^\wedge(\sigma_{\mu,\beta}))_{t,\ell} \sigma_{\mu,\beta}(\pi_\lambda(k_0\phi, k))_{\ell,t}.
\end{aligned}$$

By (3.2)(iii), for each $\varepsilon > 0$ we may choose a finite subset P_0 of P, independent of λ, such that

$$\begin{aligned}(6)\qquad &\left|(\kappa_\delta * (\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1})(\pi_\lambda(k_0\phi, k))\right. \\ &\left. - \sum_{\beta\in P} d_{\mu,\beta} \sum_{t,\ell} (\hat{\kappa}_\delta(\sigma_{\mu,\beta})((\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1})\hat{}\,(\sigma_{\mu,\beta}))(\pi_\lambda(k_0\phi, k))_{\ell,t}\right| < \varepsilon.\end{aligned}$$

Now, Lemmas 3.5 and 3.6 allow us to approximate $((\rho_{\mu,\eta})_{r,s} \circ \pi_\lambda^{-1})\hat{}\,(\sigma_{\mu,\beta})_{t,\ell}$ to within $0(1/\lambda)$, uniformly for $\beta \in P_0$, by

$$(7)\qquad K_{\mu,\lambda}^{r,s,t,\ell}\left(\eta, \frac{\beta}{\lambda}\right),$$

and $\sigma_{\mu,\beta}(\pi_\lambda(k_0\phi, k))_{\ell,t}$ to within $0(1/\lambda)$, uniformly for $\beta \in P_0$, by

$$(8)\qquad (\rho_{\mu,\beta/\lambda}(k_0\phi, R))_{\ell,t}.$$

Finally, from [1, (3.3)] and Lemma 3.5,

$$(9)\qquad |\hat{\kappa}_\delta(\sigma_{\mu,\beta})_{t,u} - \lambda^{\dim V}(\kappa \circ \pi_\lambda^{-1})\hat{}\,(\rho_{\mu,\beta/\lambda})_{t,u}| < 0\left(\frac{1}{\lambda}\right)$$

for $1 \leqq t, u \leqq M$.

Combining (6) with (5), (7), (8), and (9), we have

$$\begin{aligned}(10)\qquad &\left|(\rho_{\mu,\eta})_{r,s}(k_0\phi, k) - \sum_{\beta\in P_0} d_{\mu,\beta} \sum_{u,t,\ell=0}^{\min(M,d_{\mu,\beta})} \lambda^{\dim V}(\kappa \circ \pi_\lambda^{-1})\hat{}\,(\rho_{\mu,\beta/\lambda})_{t,u}\right. \\ &\left. \cdot K^{r,s,u,\ell}\left(\eta, \frac{\beta}{\lambda}\right)(\rho_{\mu,\eta})_{\ell,t}(k_0\phi, k)\right| < 0\left(\frac{1}{\lambda}\right) + \varepsilon.\end{aligned}$$

By (3.2)(iv), we also have

$$|\lambda^{\dim V}(\kappa \circ \pi_\lambda^{-1})\hat{}\,(\rho_{\mu,\beta/\lambda})_{t,u}| = 0\left(\frac{1}{\lambda}\right).$$

This fact, together with (3) and (10) gives

$$\left|f_\mu(x) - \sum_{\beta\in P_0} d_{\mu,\beta} \sum_{r,s,t,\ell=0}^{\min(M,d_{\mu,\beta})} K^{r,s,t,\ell}\left(\eta, \frac{\beta}{\lambda}\right) \int \hat{f}_{s,r}(\rho_{\mu,\phi})(\rho_{\mu,\phi})_{\ell,t}\left(k_0\frac{\beta}{\lambda}, k\right)\right| < 0\left(\frac{1}{\lambda}\right) + \varepsilon.$$

So, given $\varepsilon > 0$ we may choose λ_0 such that $\lambda > \lambda_0$

$$\left|f_\mu(x) - \sum_{\beta\in P_0} d_{\mu,\beta} \sum_{r,s,t,\ell=0}^{\min(M,d_{\mu,\beta})} K^{r,s,t,\ell}\left(\eta, \frac{\beta}{\lambda}\right) f_{s,r,\ell,t}\left(k_0\frac{\beta}{\lambda}, k\right)\right| < \varepsilon + 0\left(\frac{1}{\lambda}\right).$$

The assumptions on f guarantee that f is rapidly decreasing. Hence for each s, r, t, and ℓ the sum

$$\sum_{\beta\in P} d_{\mu,\beta} K^{r,s,t,\ell}\left(\eta, \frac{\beta}{\lambda}\right) f_{s,r,\ell,t}\left(k_0\frac{\beta}{\lambda}, k\right)$$

is finite. From this it follows, since ε is arbitrary, that

$$\left|f_\mu(x) - \sum_{\beta\in P} d_{\mu,\beta} \sum_{r,s,t,\ell=0}^{\min(M,d_{\mu,\beta})} K^{r,s,t,\ell}\left(\eta, \frac{\beta}{\lambda}\right) f_{s,r,\ell,t}\left(k_0\frac{\beta}{\lambda}, k\right)\right| < 0\left(\frac{1}{\lambda}\right). \qquad \square$$

4. Corollaries and complements. To obtain a more precise analogue of the sampling theorem, we would need also

$$K_{\mu,\lambda}^{r,s,t,\ell}\left(\eta, \frac{\beta}{\lambda}\right) = K_{\mu,\lambda}\left(\eta, \frac{\beta}{\lambda}\right)\delta_{r,t}\delta_{s,\ell} + 0\left(\frac{1}{\lambda}\right),$$

where $K_{\mu,\lambda}$ is independent of r, s, t, and ℓ, and δ denotes the Dirac delta. This would yield (for $x = (k_0\eta, k)$)

$$f_\mu(x) = \sum_{\beta \in P_+} d_{\mu,\beta} K_\mu\left(\eta, \frac{\beta}{\lambda}\right) f_\mu\left(k_0 \frac{\beta}{\lambda}, k\right).$$

Unfortunately, such a condition does not hold for $K^{r,s,t,\ell}$. Nevertheless, $K^{r,s,t,\ell}$ is "approximately" diagonal. Consider the right action of K on $(\rho_{\mu,\phi})_{r,s}$. We have

$$\begin{aligned}(\rho_{\mu,\phi})_{r,s}(xk) &= (\rho_{\mu,\phi}(xk)u_r, u_s)\\ &= (\rho_{\mu,\phi}(x)u_r(k^{-1}\cdot), u_s(\cdot)).\end{aligned}$$

It follows that $K^{r,s,t,\ell} = 0$ unless u_r and u_t have the same K-type. It is similar for u_s and u_ℓ. Our choice of basis thus implies that there is $n(i)$ such that r and t (respectively, s and ℓ) both lie between $n(i)+1$ and $n(i+1)$.

In the special case $(G, K) = (SO(3)), SO(2))$, we have $n(i) = i$, and thus K_μ is diagonal.

COROLLARY 4.1. *For the Euclidean motion group $M(2)$, the sampling theorem takes the form*

$$f(x) = \sum_{n \in N} (2n+1) \sum_{r=1}^{M} K^r\left(\eta, \frac{n}{\lambda}\right) f_+\left(k_0 \frac{n}{\lambda}, k\right) + 0\left(\frac{1}{\lambda}\right)$$

where

$$f_r(x) = \int_0^R (\hat{f}(S)\rho_S(x))_{r,r} S\, dS$$

and

$$K^r(\eta, \phi) = \frac{1}{\lambda^2}\int_0^\lambda J_r(S\eta)J_r(S\phi)S\, dS.$$

(*As usual, the J's denote Bessel functions.*)

By making certain restrictions on the function f we may also obtain simpler forms of the sampling theorem. In particular, if f is bi-K-invariant $\hat{f}$ has rank one: we may suppose that $\hat{f}(\rho)_{r,s} = 0$ unless $r = s = 0$. For this case $u_r = u_s = 1$, the identity on K;

$$\rho_{\mu,\phi}(v, k)_{00} = \int e^{i\phi(k^{-1}v)}\, dk = J_\phi(\eta), \quad \text{where } v = k_0\eta.$$

The argument on K-types given above shows that $K_{\mu,\lambda}^{r,s,t,\ell} = 0$ unless $\mu = 1$ and $r = s = t = \ell = 0$. Thus we see that

$$K_\lambda = K_\lambda^{00} = \lambda^{\dim V}\int_{\beta_\lambda} J_0(\eta)\overline{J_0\left(\frac{\beta}{\lambda}\right)} \prod_{\alpha \in P_+} \phi(H_\alpha)\, d\phi.$$

THEOREM 4.2. *Let f be bi-K-invariant on G. Then*

$$\left| f(x) - \sum_{\beta \in P_+} d_{1,\beta} K_\lambda\left(\eta, \frac{\beta}{\lambda}\right) f\left(k_0 \frac{\beta}{\lambda}, k\right)\right| < 0\left(\frac{1}{\lambda}\right).$$

COROLLARY 4.3. *Let f be a K-invariant function on V. Then*

$$\left| f(\eta) - \sum_{\beta \in P_+} d_\beta K_\lambda\left(\eta, \frac{\beta}{\lambda}\right) f\left(\frac{\beta}{\lambda}\right) \right| < 0\left(\frac{1}{\lambda}\right).$$

With this corollary, taking $(G, K) = (SO(n+1), SO(n))$, we recover the case of even spacing for the n-Hankel transform on R.

5. Applications. In this section we should like to mention two possible applications of the above theorems in the area of image processing and signal analysis. We will restrict ourselves to exploitation of Corollaries 4.1 and 4.3, although it is clear that the method would work more generally. Even in these special cases, work remains to be done to implement the sampling in any practical situation. However, we believe they indicate the relevance of the theory presented here to the real world.

Example 5.1. Image reconstruction in the plane. In many applications, there is interest in reconstructing a two-dimensional picture from a two-dimensional set of data points that are equally spaced along radial lines at angles $\theta_1 \cdot \theta_2 \cdots, \sigma_k$, for instance. As examples, we mention problems in geophysics [10], in radiology (the fan beam problem—see [3]) and in radar (see [11]).

Corollary 4.3 may be applied in this context to give an approximate reconstruction technique: if the points are spaced at distance $1/\lambda$ apart, and if the function f is Fourier-Bessel band limited, then in polar coordinates, we have

$$f(\eta, \theta) = \sum_{n \in N} (2n+1) \sum_{|r| \leqq M} K^r\left(\eta, \frac{n}{\lambda}\right) f_r\left(\frac{n}{\lambda}, \theta\right) + 0\left(\frac{1}{\lambda}\right),$$

where $f_r(\eta, \theta) = \int f(\eta, \phi)\, e^{-ir\phi}\, d\phi \cdot e^{ir\sigma}$, and K^r is as in (4.1). Standard Shannon sampling may then be applied in the angular variables, giving

$$f(\eta, \theta) = \sum_{n \in N} \sum_{k} (2n+1) \sum_{|r| \leqq M} K^r\left(\eta, \frac{n}{\lambda}\right) e^{-i(\theta - \theta_k)} f_r\left(\frac{n}{\lambda}, \theta_k\right) + 0\left(\frac{1}{\lambda}\right).$$

This represents approximate reconstruction of f from its values at the points $(n/\lambda, \theta_k)$.

Example 5.2. Digital signal processing. Another scenario in which our results may be applied is that of a one-dimensional signal of which the n-Hankel transform is known to be band limited, and which is being sampled at equally spaced time intervals. This typically arises when signal sampling is done by a digital processor. Corollary 4.3 then enables us to write

$$f(x) = \sum_{k=0}^{\infty} d_k K_\lambda\left(x, \frac{k}{\lambda}\right) f\left(\frac{k}{\lambda}\right) + 0\left(\frac{1}{\lambda}\right),$$

where $d_k = (2k+n-2)(n+k-3)!/(n-2)!k!$ is the dimension of the space of harmonic polynomials of degree k in R^n and where K_λ is as in Corollary 4.3. Again, this shows how to recover f to within $0(1/\lambda)$ from evenly spaced data points.

Acknowledgment. The idea for this paper occurred to me during a seminar given by I. Kluvánek. I gratefully acknowledge this source of my inspiration.

REFERENCES

[1] A. H. DOOLEY AND G. I. GAUDRY, *On L^p multipliers of Cartan motion groups*, Journal of Functional Analysis, 67 (1986), pp. 1-24.

[2] A. H. DOOLEY AND J. W. RICE, *On contractions of semisimple Lie groups*, Trans. Amer. Math. Soc., 289 (1985), pp. 185-202.

[3] W. G. HAWKINS AND H. H. BARRET, *A numerically stable circular harmonic reconstruction algorithm*, SIAM J. Numer. Anal., 23 (1986), pp. 873-890.

[4] S. HELGASON, *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York, 1978.

[5] E. HEWITT AND K. A. ROSS, *Abstract Harmonic Analysis*, Vol. II, Springer-Verlag, Berlin, New York, 1970.

[6] A. JERRI, *The Shannon sampling theorem: A tutorial review*, Proc. IEEE, 65 (1977), pp. 1565-1596.

[7] ———, *Sampling expansion for the L_ν^α-Laguerre integral transform*, J. Research of Nat. Bur. Stands-B Math. Sci, 80B (1976), pp. 415-418.

[8] A. KLEPPNER AND R. L. LIPSMAN, *The Plancherel Formula for group extensions I*, Ann. Sci. Ecole Norm. Sup., 5 (1972), pp. 459-516.

[9] H. P. KRAMER, *A generalized sampling theorem*, J. Math. Phys., 38 (1959), pp. 68-72.

[10] R. G. ROGERS AND S. EDWARDS, *Iterated projection and extrapolation of Radon transform data using the Paley-Wiener theorem*, to appear.

[11] W. SCHEMPP, *Analog radar signal design and digital signal processing—a Heisenberg nilpotent Lie group approach*, in Lie Methods In Optics, J. Sánchez-Mandragon and K. B. Wolfe, eds., Springer Lecture Notes in Physics 250 (1986).

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 634-642, May 1989

010

ON SOME CONJECTURES OF TURCOTTE, SPENCE, BAU, AND HOLMES*

S. P. HASTINGS† AND W. C. TROY†

Abstract. Turcotte, Spence, and Bau [*Internat. J. Heat Mass Transfer*, 25 (1982), pp. 699-706] contains conjectures concerning the equation

$$V'' = (V^2 - A(1-X^2))/2$$

with boundary conditions $V(-1) = V(1) = 0$, where A is a nonnegative parameter. For large A an appropriate asymptotic expansion results in a version of the first Painlevé transcendent, namely $Y'' = (Y^2 - s)/2$ seeking a solution such that $Y(0) = 0$, $Y(s) \sim \sqrt{s}$ as $s \to \infty$. This was studied extensively by Holmes and Spence, who conjectured that there are only two solutions. In this paper proofs of these conjectures are provided. During one of these proofs it is shown how a computer language for symbol manipulation, such as MACSYMA, can be used in a mathematically rigorous analysis.

Key words. boundary value problems, MACSYMA

AMS(MOS) subject classification. 34B15

1. Introduction and statement of results. In [2] Turcotte, Spence, and Bau have studied the problem of vertical flow of an internally heated Boussinesq fluid with viscous dissipation and pressure work. For the steady state flow under appropriate assumptions they have obtained the following boundary value problem for the velocity V as a function of the scaled position x, where the walls of the vertical channel are at $x = \pm 1$:

$$V''(x) = (V(x)^2 - A(1-x^2))/2, \tag{1}$$

$$V(1) = V(-1) = 0. \tag{2}$$

Here A is a nonnegative parameter. They have studied separately the cases $A = 0$, A small, A intermediate, and A large. In the latter cases their numerical observations have led them to conjecture that as A increases the number of solutions of (1)-(2) increases without bound.

In studying the case A large, they have found it necessary to match an inner expansion near the walls $x = \pm 1$ with an outer solution valid in the interior. (The outer solution is simply $V = \pm\sqrt{A(1-x^2)}$.) They have found the appropriate inner variables near $x = -1$ to be

$$s = (1+x)/\varepsilon, \qquad Y = \varepsilon^2 V$$

where $\varepsilon = (2A)^{-1/5}$. In terms of these variables the lowest-order terms of the inner expansion have been found to satisfy the equation

$$Y'' = (Y^2 - s)/2 \tag{3}$$

with boundary conditions

$$Y(0) = 0, \qquad Y(s) \sim \sqrt{s} \quad \text{as } s \to \infty. \tag{4}$$

This problem has been, in turn, studied extensively by Holmes and Spence [1]. They have obtained several interesting results, which are summarized in § 3. However,

* Received by the editors January 25, 1988; accepted for publication (in revised form) June 16, 1988. This research was supported by the National Science Foundation.

† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

they left unanswered a key question that we address in this paper, namely, the number of solutions of (3)-(4) that match the outer solution $Y=+\sqrt{s}$ at $+\infty$. We show that there are exactly two such solutions. This had been conjectured by Holmes and Spence on the basis of extensive numerical calculations. We give two proofs of this result. One relies on detailed (four decimal place) estimates of the solution while the other uses a symbolic manipulator (MACSYMA) to determine the number of zeros of certain comparison polynomials and thereby implement the previous method with less effort. The steps using MACSYMA are discussed throughout the more standard proof. Note that the result using MACSYMA is rigorous, because all computations using rational arithmetic are exact. We make crucial use of Sturm's theorem on the number of roots of a polynomial in a given interval [3]. We also investigate the behavior of solutions of the original problem (1)-(2). In Theorem 1 we prove the conjecture made by Turcotte, Spence, and Bau.

Our results are the following theorems.

THEOREM 1. *Let $N(A)$ denote the number of solutions of* (1)-(2). *Then $N(A)\to\infty$ as $A\to\infty$.*

THEOREM 2. *The differential equation* (3) *has exactly two solutions such that $Y(0)=0$ and $Y(s)\sim\sqrt{s}$ as $s\to\infty$. For one of these solutions, say Y_+, $Y'(0)>0$, while for the other, say Y_-, $Y'(0)<0$.*

See [1] and [2] for the implications of these results.

2. Proof of Theorem 1. This is by far the easier of the two results to prove. We will show, in fact, that there are, for large A, many solutions of (1) that are even ($V'(0)=0$). The proof begins by describing the behavior of a particular solution of (1). Let $\hat{V}$ denote the unique solution of (1) such that $V'(0)=0$ and $V(0)=-\sqrt{A}$.

LEMMA 1. *The solution $\hat{V}$ increases monotonically to infinity at a finite blowup point $x_A>1$. Further, $\hat{V}''>0$ on $(0, x_A)$.*

Proof. We consider the energy functional

$$E(x)=\frac{(\hat{V}''')^2}{2}-\frac{\hat{V}(\hat{V}'')^2}{2}-\hat{V}''(\hat{V}')^2-A\hat{V}''$$

so that

$$E'(x)=-\frac{5\hat{V}'(\hat{V}'')^2}{2}$$

and E decreases as long as $\hat{V}'$ is positive. Note that $E(0)=0$. If there is a first point where $\hat{V}''=0$, then at that point E is positive, which is a contradiction. The rest of the conclusions follow readily.

We now use a "shooting" method to obtain solutions of (1)-(2). It seems best to "shoot" backward from $x=1$. Therefore let V_β denote the unique solution of (1) such that $V(1)=0$ and $V'(1)=\beta$. We will show that for large A there are many values of β such that $V'_\beta(0)=0$, which will prove Theorem 1.

LEMMA 2. *If $\beta=0$, then V_β is negative on* $[0, 1)$.

Proof. In this proof, V denotes V_0. By differentiating (1) it is seen that $V'''(1)=A>0$, so that in some interval to the left of 1, $V<0$. Suppose that there is a first positive x_1 to the left of 1 where $V=0$. Let F be the energy function defined by

$$F=(V')^2-\frac{V^3}{3}+AV-Ax^2V.$$

Then

$$F' = -2AxV$$

so that as x decreases from 1, F decreases from zero as long as V is negative. However, at x_1, $F \geqq 0$, which is the desired contradiction.

Now let $U(x) = \hat{V}(x) - V_0(x)$. Then

$$U''(x) = \tfrac{1}{2}(\hat{V} + V_0)U(x)$$

so that U oscillates frequently when $\hat{V} + V_0$ is large and negative. It follows from Lemma 1 that $\hat{V} \leqq -\sqrt{A(1-x^2)}$ on $[0, 1]$. Thus, on $[0, \frac{1}{2}]$, say, $V_0 + \hat{V}$ is large and negative for large A. Hence U must vanish many times in this interval, for large A.

On the other hand, it is easy to show that if β is large (for fixed A), then $\hat{V} - V_\beta$ vanishes only once in $[0, 1)$, at a point close to $x = 1$. The graph of V_β cannot be tangent to the graph of $\hat{V}$, and it is then seen that the zeros of $\hat{V} - V_\beta$ that exist when $\beta = 0$ must leave the interval $[0, 1)$ by crossing to the left across $x = 0$. As successive zeros cross $x = 0$, $V'_\beta(0)$ must alternate in sign, since $\hat{V}'(0) = 0$. Therefore, between values of β such that $V_\beta(0) = 0$ there must be values of β such that $V'_\beta(0) = 0$, and these correspond to solutions of (1)-(2). This proves Theorem 1.

3. Proof of Theorem 2. (i) It is convenient to rescale the equation, setting $y(x) = aY(bx)$, with $a = 4^{1/5}$ and $b = a^2/2$. This leads to the equation

$$y'' = y^2 - x. \tag{5}$$

We want a solution such that

$$y(0) = 0, \qquad y \sim \sqrt{x} \quad \text{as } x \to \infty. \tag{6}$$

As before, we use a shooting method. Let y_γ denote the solution of (5) such that

$$y(0) = 0, \qquad y'(0) = \gamma.$$

Holmes and Spence [1] have proved a unique positive γ exists such that y_γ satisfies (5)-(6), and there is at least one such value of γ that is negative. They show that the solution with $\gamma = 0$ does not satisfy (3). They have conjectured that there is only one possible negative γ. Holmes and Spence have further proved that for each negative γ, there is a first $x = x_\gamma > 0$ where $y'_\gamma(x) = 0$. Finally, they have proved the following important result.

LEMMA 3. *Suppose that in some interval $\gamma_1 < \gamma \leqq \gamma_2 < 0$, x_γ is a strictly decreasing function of γ, and also that $y_\gamma(x_\gamma)$ is decreasing in γ. Then there is at most one value of γ in this interval such that y_γ satisfies* (6).

Recall that Holmes and Spence have shown at least one γ^* exists such that (6) is satisfied if $\gamma = \gamma^*$. We prove Theorem 2 with the following two additional lemmas.

LEMMA 4. *Any possible γ^* is less than -2.3.*

LEMMA 5. *The hypotheses of Lemma 3 hold with $\gamma_1 = -\infty$ and $\gamma_2 = -2.3$.*

Proof of Lemma 4. We use the functional

$$Q = \frac{y^3}{3} - xy - \frac{(y')^2}{2},$$

where $y(x)$ is a solution of (5)-(6). Thus, $Q(0) = -\gamma^2/2$ and

$$Q' = -y. \tag{7}$$

Note that, on some interval $(0, \varepsilon)$, $Q < 0$ and $y < 0$. The following useful result is easy to prove.

LEMMA 6. *If $Q>0$ before $y>0$, then y cannot satisfy* (6).

Proof. If (6) holds then there is a first x_1 with $y(x_1)=0$. From (7) we see that $Q(x_1)>0$. However, the definition of Q shows that $Q(x_1)\leqq 0$. This proves Lemma 6.

Continuing with the proof of Lemma 4, our method requires step-by-step estimates of the solution, via a sequence of further lemmas.

LEMMA 7. *For $0<x<.6$, as long as $y'<0$, y'' has at most one zero.*

Proof of Lemma 7. Consider the functional

$$H=(y''')^2/2-y(y'')^2-2y''(y')^2.$$

Then $H(0)=\frac{1}{2}$ and

$$H'=-5y'(y'')^2. \tag{8}$$

If there are at least two numbers $b>a>0$ in $(0, .6)$ where $y''=0$, and $y'<0$ in $[0, b)$, then (8) implies that $H\geqq\frac{1}{2}$ in $[0, b]$. Therefore, from the definition of H, $y'''\neq 0$ at $x=a$ and $x=b$. We may therefore assume that $y''<0$ on $(0, a)$ and $y''>0$ on (a, b). From (8) and the definition of H we conclude that

$$y'''(a)\geqq 1, \tag{9}$$

$$y'''(b)\leqq -1. \tag{10}$$

Differentiating (5), we obtain

$$y''''=2yy''+2(y')^2\geqq 2y^3.$$

On (a, b), $|y|<\sqrt{b}$, so $y''''\geqq -2(.6)^{3/2}$. Integrating this and using (9) gives a contradiction to (10). This proves Lemma 7.

Continuing with the proof of Lemma 4, we need Lemma 8.

LEMMA 8. *Suppose that $\gamma<0$. If y'' has two zeros on $(0, x_\gamma]$, where x_γ is the first zero of y', then y cannot satisfy* (6).

Proof. Again let a and b be the first two zeros of y'', and assume that $y'<0$ in $[0, b)$. From the previous lemma we see that $b>.6$. Also, as before, $y'''(a)>1$, $y'''(b)<-1$, and $y''>0$ on (a, b). It follows that $y(b)=-\sqrt{b}$ and $y'(b)\geqq-\frac{1}{2}\sqrt{b}$. Therefore, $Q(b)\geqq 2b^{2/3}/3-1/(8b)>0$ and Lemma 6 implies the result.

As we continue the proof of Lemma 4 with the use of Lemma 6, we introduce a key method of polynomial estimates of the solution. This method can be implemented with MACSYMA. It is convenient to introduce operators on the space of polynomials as follows:

$$I(p)(x)=\int_0^x p(s)\,ds,$$

$$J(p)(x)=\gamma+I(p)(x), \qquad K(p)(x)=(p(x))^2-x,$$

$$M(p)=IJK(p)=\gamma x+\int_0^x\int_0^t (p^2(t)-t)\,dt\,dx.$$

We then define a sequence of polynomials $y_0(x)=0$, $y_{i+1}=M(y_i)$. As long as $y_2(x)$ is negative, we have

$$0>y_{2j}(x)>y_{2j+2}(x)>y(x)>y_{2j+1}(x)>y_{2j-1}(x), \tag{11a}$$

$$0>y'_{2j}(x)>y'_{2j+2}(x)>y'(x)>y'_{2j+1}(x)>y'_{2j-1}(x) \tag{11b}$$

for $j=1, 2, 3, \cdots$. The repeated use of these inequalities will give our result.

The "hand" computation first uses y_2. Easy estimates show that $y_2(x)<0$ for $0<x\leqq 1.75$ and $-2.3\leqq\gamma<0$. At this point, the "hand" computation requires that we break up the interval $[-2.3, 0)$ into $[-2.3, -2.1)$ and $[-2.1, 0)$. Using y_2 and (7), we

show that $Q(1.75)>0$ if $-2.1\leqq\gamma<0$. In the range $-2.3\leqq\gamma\leqq-2.1$, the estimate using y_2 is inadequate. However, it is difficult to work with y_4, since it involves terms as high as x^{38}. Instead, we observe that y_2 can be truncated. We find that the inequalities $y<y_2\leqq 0$ imply

$$y\leqq\bar{y}_2=\frac{\gamma^2x^4}{12}+\frac{89}{90}\gamma x-.166x^3<0. \tag{12}$$

From this we have

$$y'>JK(\bar{y}_2)$$

and

$$y>M(\bar{y}_2). \tag{13}$$

We substitute $x=1$ in these inequalities to obtain

$$y(1)\geqq-2.062,\qquad y'(1)\geqq-1.28 \tag{14}$$

if $-2.3\leqq\gamma\leqq-2.1$ and $y_2\leqq 0$ on $(0,1)$.

Next, truncating (13) gives

$$y(x)>\bar{y}_3(x)=.07\gamma^2x^4-x^3/6+\gamma x$$

for x in $[0,1]$ and γ in $[-2.5,-2.1]$. From this,

$$y\leqq M(\bar{y}_3)$$

and this inequality implies that

$$y(1)\leqq-1.9,\qquad y'(1)\leqq-1.1. \tag{15}$$

Finally, using (13) and the fact that $Q'=-y$, we obtain that

$$Q(1)\geqq-1.55 \tag{16}$$

for γ in $[-2.3,-2.1]$. Again, all of this derives from the observation that $y_2<0$ on $[0,1)$.

By this point the reader will have a clear idea of our method—repeated application of the inequalities (11) with appropriate truncations to make the calculations tractable. Our goal is still to prove that Q becomes positive before y, if $\gamma\geqq-2.3$. Continuing on from $x=1$, with (14)-(16) as starting values, accomplishes this. At the same time, it can be verified that $y_2<0$ as long as necessary. At this point, however, we turn to a proof using MACSYMA.

4. Proof of Lemma 4 using MACSYMA. We can, of course, simply use MACSYMA to check the various steps of the earlier proof, and indeed, this has been done. However, we can better illustrate the utility of symbolic manipulation as a tool for this sort of proof by an application that does not need the nontrivial truncation that obtaining (12) required. In fact, it turns out that Lemma 4 can be strengthened, and this improvement helps give a complete MACSYMA proof of Theorem 2.

LEMMA 4M. *If* $-2.4\leqq\gamma<0$, *then* $Q=0$ *before* $y=0$, *so that* (6) *cannot hold.*

Proof. We use Sturm's theorem [3]: Let p be a polynomial of degree n. Define polynomials $p_0, p_1, \cdots, p_r$ as follows: $p_0=p$, $p_1=p'$, $p_i=q_ip_{i+1}-p_{i+2}$ where the q_i are polynomials, $\deg(p_{i+2})<\deg(p_{i+1})$, and $p_r\neq 0$, $p_{r+1}=0$ (Euclidean algorithm). Then the number of roots of p in an interval $(x_1,x_2]$, not counting multiplicity, is $j-k$, where j is the number of sign changes (ignoring zeros) in the sequence $\{p_0(x_1),p_1(x_1),\cdots,p_r(x_1)\}$ and k is the number of sign changes when the p_i are evaluated at x_2.

It is now clear how, in principle, MACSYMA can be used to give rigorously the number of roots of a polynomial p with rational coefficients in an interval with rational

endpoints, since MACSYMA calculations with rational numbers are exact. However, limitations of machine time and memory mean that all is not completely straightforward. With the VAX 8600, polynomials of degree 15 or 16 can generally be handled within 10 minutes of CPU time (although this depends on how many digits are required to give the coefficients as fractions), but machine time and space requirements quickly become a problem for polynomials of higher degree.

Further, this theorem is for a polynomial in one variable, and our comparison polynomials y_j depend on x and γ. Therefore some investigation is necessary. The hand calculations made earlier facilitate this. We first prove Lemma 9.

LEMMA 9. *If* $0 \leqq x \leqq 1.8$ *and* $-2.4 \leqq \gamma \leqq 0$, *then* $y < 0$.

Proof. We use a series of intermediate results.

LEMMA 10. *If* $-2 \leqq \gamma \leqq 0$ *and* $0 \leqq x \leqq 1.8$, *then* $y_2 < 0$.

Proof. Since y_2 is of degree 8 in x, this can be checked with Sturm's theorem (computations done with MACSYMA) for $\gamma = 0$ and $\gamma = -2$. Since $d^2y_2/d\gamma^2$ is positive, y_2 must also be negative for the intermediate values of γ.

Now let I and J denote the intervals $(1.7, 1.8]$ and $[-2.4, -2]$, respectively.

LEMMA 11. $y_2' > 0$ *in* $I \times J$.

Proof. $y_2'(1.7, \gamma)$ is a quadratic polynomial in γ and it is easy to check that it is positive for $\gamma \in J$. Also, $y_2'' = y_1^2 - x$. But $y_1'' = -x < 0$, while $d^2/dx^2(-\sqrt{x}) > 0$. Again, we easily check that $(y_1^2 - x) > 0$ in $I \times J$.

LEMMA 12. $y_3' < 0$ *in* $I \times J$.

Proof. $y_3'(1.7, \gamma)$ is of degree 4 in γ. With MACSYMA this is easily found to be negative for γ in J. Also, $y_3'' = y_2^2 - x$. Further use of Sturm's theorem shows that for γ in I, $|y_2(x, \gamma)|^2 < 1.8$ for $x = 1.7$ and $x = 1.8$. Hence, from Lemma 11, $y_3'' < 0$ in $I \times J$. This implies Lemma 12.

LEMMA 13. $y_3^2 > x$ *in* $I \times J$.

Proof. The polynomial $y_3^2(1.7, \gamma)$ is of degree 8 in γ, so Sturm's theorem can be used to show that $y_3^2(1.7, \gamma) > 1.8$. Now Lemma 13 follows from Lemma 12.

LEMMA 14. $y_4' > 0$ *in* $I \times J$.

Proof. Similarly, $y_4'(1.7, \gamma)$ is of degree 8 in γ, and, using MACSYMA, we easily show it is positive. Lemma 14 therefore follows from Lemma 13.

To prove Lemma 9, we first use MACSYMA to check that y_2 is negative for x in $(0, 1.7)$ and γ in $[-2.4, 0]$. Hence, by (11), $y < y_4 < y_2 < 0$ in this region. If $-2 \leqq \gamma \leqq 0$, then these inequalities hold up to $x = 1.8$. For γ in J, we observe that $y < y_4$ as long as $y_4^2 > y_2^2$. Since y_2 is increasing in I, the last inequality can only fail if, at some x in I, $y_4 + y_2 = 0$. As a result of Lemmas 11 and 14, to prove Lemma 9 we need only check that $y_4 + y_2 \leqq 0$ at $x = 1.8$, γ in J. This polynomial, of degree 8 in γ, is again easily checked. This completes the proof of Lemma 9.

Lemma 4M is a consequence of Lemmas 6 and 9 and the following lemma.

LEMMA 15. *If* $0 < -\gamma \leqq 2.4$, *then* $Q(1.8) > 0$.

Proof. From (11) and (7),

$$Q(x) \geqq -\frac{\gamma^2}{2} - \int_0^x y_4(s)\, ds$$

for (x, γ) in the region described in Lemma 9.

Setting $x = \frac{9}{5}$ and using Sturm's lemma on $-\frac{12}{5} < \gamma \leqq 0$ gives the result.

5. Proof of Theorem 2. (ii) The idea here is to use the variational equation for (5). This is

$$w'' = 2yw. \tag{17}$$

To show that the hypotheses of Lemma 3 are satisfied in $(-\infty, -2.3)$, which is necessary for the "hand" proof, we must show that x_γ, the first zero of y'_γ, and $y_\gamma(x_\gamma)$ are strictly increasing functions of x_γ in an appropriate interval.

It is convenient to use the notation $y_\gamma(x) = y(x, \gamma)$ in this section. Recall that γ^* denotes some value of γ such that (6) is satisfied. We show first that $y(x_\gamma, \gamma)$ is strictly increasing in $(-\infty, \gamma^*)$.

Note that from the definition of x_γ,

$$\frac{d}{d\gamma}[y(x_\gamma, \gamma)] = \frac{\partial y(x, \gamma)}{\partial \gamma}\bigg|_{x=x_\gamma} = w(x_\gamma, \gamma),$$

where $w(x, \gamma)$ is the solution of (17) with $y = y(x, \gamma)$ and $w(0) = 0$, $w'(0) = 1$. The Sturm Oscillation Theorem implies that $w(x, \gamma) > 0$ for $0 < x \leqq x_\gamma$ if

$$-2x_\gamma^2 y(x_\gamma, \gamma) < \pi^2. \tag{18}$$

The "hand" proof of (18) begins with the relations

$$\bar{y}_3(x) = .07\gamma^2 x^4 - \frac{x^3}{6} + \gamma x < y < \bar{y}_2 = \frac{\gamma^2 x^4}{12} + \frac{89\gamma x}{90} - (.166)x^3, \tag{19}$$

which hold as long as $\bar{y}_2 < 0$.

The key step is to find upper and lower bounds in terms of γ for x_γ. Numerical and asymptotic calculations suggest that these bounds should each be of the form $-k/\gamma^{1/3}$ for some k, and numerical experimentation suggests the values 1.32 and 1.85 for k. Therefore we define an initial value $x_0 = -1.32/\gamma^{1/3}$, and set $x_1 = -1.85/\gamma^{1/3}$. Substituting x_0 into (19) and truncating appropriately, we obtain the relations

$$(.3863)\gamma - \frac{.64}{\gamma^{2/3}} - \frac{.0274}{\gamma^{7/3}} \leqq y'(x_0) \leqq (.34)\gamma - \frac{.63}{\gamma^{2/3}} - \frac{.028}{\gamma^{2/3}} \tag{20}$$

and

$$-(1.099)\gamma^{2/3} + \frac{.331}{\gamma} + \frac{.0045}{\gamma^{8/3}} \leqq y(x_0) \leqq -(1.0894)\gamma^{2/3} + \frac{.327}{\gamma}. \tag{21}$$

We also see that $\bar{y}_2(x_1) < 0$.

To prove (18), we first show that $y'' > 0$ on (x_0, x_γ). This implies that

$$y(x_\gamma) > y(x_0) + y'(x_0)(x_1 - x_0),$$

and substitution of the estimates for x_1 and x_0 into $x_\gamma^2 y(x_\gamma)$ yields (18).

6. Proof of (18) using MACSYMA. Our goal is again to prove (18), and the technique is the same as before; that is, we find upper and lower bounds for x_γ, the first zero of y', and then use the comparison polynomials to estimate $x^2 y(x)$ at x_γ.

A minor complication in this section is the need to consider all $\gamma \leqq -2.4$ (from Lemma 4M). Sturm's lemma requires consideration of polynomials in γ over a finite interval. However, we can circumvent this by using MACSYMA to prove the following lemma.

LEMMA 16. *If $\gamma \leqq -100$, then y cannot satisfy* (6).

Proof. We can readily see, even by hand, that $y_3(1) > 0$ if $\gamma < -100$. Thus the first zero of y must come before $x = 1$. It is easy to show, further, that $\min_{0 \leqq x \leqq 1} y(x) \to -\infty$ as $\gamma \to -\infty$. From this we quickly see that $y'(x_1) \to \infty$, and the result follows.

Nevertheless, implementing Sturm's lemma was a little harder than expected and proved to be a valuable lesson in using these techniques. By Lemmas 4M and 16, we must obtain the desired estimate for $-100 \leqq \gamma \leqq -2.4$. It turns out that if we could

obtain the estimate at $\gamma = -2.4$, then using Sturm's lemma gives the same estimate in $-100 < \gamma \leqq -12/5$.

Initially, then, we tried to demonstrate (18) for $\gamma = \gamma_0 = -12/5$. Using y_2' we first obtain a lower bound on x_{γ_0}. As before, we look for this estimate in the form $-k/\delta_0$, where $\delta_0 = \gamma_0^{1/3}$. From y_2' we find that $k = 7/5$ is sufficient. That is, y_2', and hence y', are negative at $x = -7/(5\delta)$ if $-100^{1/3} < \delta \leqq \delta_0$. Also, $y_3'' > 0$ for $x > -7/(5\delta)$, which implies that $y'' > 0$ as long after $-7/(5\delta)$ as $y_2 < 0$.

To obtain an upper bound for x_γ, we must use y_3', looking for k_1 such that $y_2(-k_1/\delta) < 0$ and $y_3'(-k_1/\delta) > 0$, if $\delta \leqq \delta_0$. We found that $k_1 = 191/100$ satisfied these conditions. Unfortunately, these estimates are insufficient. We find that if $x_0 = -7/(5\delta_0)$ and $x_1 = -191/(100\delta_0)$, then

$$-x_0^2(y_3(x_0) + y_3'(x_0)(x_1 - x_0)) > \pi^2/2. \tag{22}$$

In other words, the estimates based on y_2 and y_3 are insufficient, and we must use the 38th-degree polynomial y_4. In principle, this is not difficult, because we need to use y_4 only to improve the estimate of x_0, and we can easily show that $y_2(x_1) < 0$. In practice, however, it seemed we would be unable to implement the method because to verify (18) on an interval (γ_1, γ_0), we would need to apply the Euclidean algorithm to the polynomial $P(\delta) = -\delta^{37} y_4'(-k/\delta)$, where k is chosen so that y_4', and hence y', is negative at $-k/\delta_0$. This polynomial is of degree 40, and in general, as observed earlier, it seemed that exactly implementing the Euclidian algorithm with rational arithmetic would exceed time and space limitations on the computer. Fortunately, however, the polynomial P, while of high degree, only contains nine terms, and the computation does succeed in a reasonable time. We find that $y_4'(-8/(5\delta))$ is negative for $-100^{1/3} < \delta \leqq \bar{\delta}_0$, where $\bar{\delta}_0 = -1{,}338/1{,}000$ is an upper bound for δ_0. Using $x_0 = -8/(5\delta)$ and $x_1 = -191/(100\delta)$ in (22) yields the required reverse inequality for $-100^{1/3} < \delta \leqq \bar{\delta}_0$.

7. Continuation of proof of Lemma 5. We have only to show that x_γ is increasing in γ for $\gamma \leqq \gamma^*$. One technical difficulty is the possibility that there are values $\gamma_2 < \gamma^*$ such that $y(x, \gamma_2)$ attains both a relative minimum and relative maximum in the interval $0 < x < x_\gamma^*$.

LEMMA 17. *Suppose that $\gamma < \gamma^*$ and as before let x_γ denote the first positive zero of $y'(x, \gamma)$. If $0 < x_\gamma < x_\gamma^*$, then $y''(x_\gamma, \gamma) > 0$. If $y'(a, \gamma) = 0$ at some first a in (x_γ, x_γ^*), then $y(a, \gamma) > y(a, \gamma^*)$.*

Proof. First, we determine some crucial properties of $y(x, \gamma^*)$. It easily follows from (5) that the first zero of $y''(x, \gamma^*)$ lies in the interval $(0, 1/\gamma^{*2}) \subseteq (0, \frac{1}{2})$. Furthermore, Lemma 8 implies that $y''(x, \gamma^*)$ cannot have a second zero on $(0, x_\gamma^*)$. By hypothesis, $y'(x_\gamma, \gamma) = 0$ and $0 < x_\gamma < x_\gamma^*$. The definition of x_γ implies that $y''(x_\gamma, \gamma) \geqq 0$. Suppose that $y''(x_\gamma, \gamma) = 0$. Then x_γ must be at least the second zero of $y''(x, \gamma)$, and Lemma 8 implies that $x_\gamma \geqq .6$. Furthermore, (18) guarantees that if $y(x_\gamma, \gamma) \geqq y(x_\gamma^*, \gamma^*)$, then $y(x_\gamma, \gamma) < y(x_\gamma, \gamma^*)$. But then $y''(x_\gamma, \gamma^*) < 0$, a contradiction. Therefore, $y''(x_\gamma, \gamma) > 0$.

We assume, for the sake of contradiction, that $y'(b, \gamma) = 0$ for some first b in (x_γ, x_γ^*) and that $y(x_\gamma, \gamma) \geqq y(b, \gamma^*)$. Then $y''(b, \gamma^*) \leqq 0$, a contradiction, which proves Lemma 12.

LEMMA 18. *Let $\gamma_2 < \gamma_1 \leqq \gamma^*$. Then $x_{\gamma_2} < x_{\gamma_1}$.*

Proof. Suppose, on the contrary, that there exist $\gamma_2 < \gamma_1 \leqq \gamma^*$ with $x_{\gamma_2} \geqq x_{\gamma_1}$. This and (18) imply that $y(x, \gamma_2) < y(x, \gamma_1) < 0$ for all x in $(0, x_{\gamma_1})$. Therefore $y(x_{\gamma_1}, \gamma_2) \leqq y(x_{\gamma_1}, \gamma_1)$ and

$$y'(x_{\gamma_1}, \gamma_2) \leqq 0. \tag{23}$$

Let $\tau = y(x, \gamma_2)y'(x, \gamma_1) - y(x, \gamma_1)y'(x, \gamma_2)$. Then

$$\tau' = (x + y(x, \gamma_2)y(x, \gamma_1))(y(x, \gamma_1) - y(x, \gamma_2)) > 0$$

for all x in $(0, x_{\gamma_1})$. Since $\tau(0) = 0$, this implies that

$$\tau(x_{\gamma_1}) > 0. \tag{24}$$

However, (23) and the definition of τ lead to

$$\tau(x_{\gamma_1}) = -y(x_{\gamma_1})y'(x_{\gamma_1}, \gamma_1) \leqq 0,$$

contradicting (24). This proves Lemma 18.

From (18) and the variational equations (17) it follows that $d/d\gamma[y(x_\gamma, \gamma)] > 0$ for all $\gamma < \gamma^*$. Therefore if $\gamma_2 < \gamma_1 < \gamma^*$ then $y(x_{\gamma_2}, \gamma_2) < y(x_{\gamma_1}, \gamma_1)$. This completes the proof of Lemma 4 and Theorem 2.

REFERENCES

[1] P. HOLMES AND D. A. SPENCE, *On a Painlevé-type boundary value problem*, Quart. J. Mech. Appl. Math., 37 (1984), pp. 525-538.

[2] D. L. TURCOTTE, D. A. SPENCE, AND H. H. BAU, *Multiple solutions for natural convective flows in an internally heated, vertical channel with viscous dissipation and pressure work*, Internat. J. Heat Mass Transfer, 25 (1982), pp. 699-706.

[3] B. L. VAN DER WAERDEN, *Modern Algebra*, Vol. I, Unger, New York, 1953, p. 220.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 643-663, May 1989

011

EQUILIBRIUM OF AN ELASTIC SPHERICAL CAP PULLED AT THE RIM*

P. PODIO-GUIDUGLI†, M. ROSATI‡, A. SCHIAFFINO§, AND V. VALENTE‡

Abstract. For thin and shallow caps the title problem is carefully formulated. The outcome is a nonlinear system of two ordinary differential equations of second order; this system is amenable to a variational format through reduction to a single functional equation, which turns out to be the Euler-Lagrange equation of a suitable energy integral depending on a load parameter π_0 and a thickness parameter κ_0.

It is shown that, for all admissible values of the parameters, a global minimizer exists that is unique for sufficiently large outward tractions; moreover, no matter what the cap's thickness, such a global minimizer tends to a flat pseudoconfiguration when $\pi_0 \to +\infty$. It is also shown that, for $\pi_0 = 0$, in addition to the unstressed reference configuration, a κ_0-sequence of local minimizers exists, interpretable as everted stressed configurations of the cap; this sequence, for $\kappa_0 \to +\infty$, tends to a pseudoconfiguration that is the reflection with respect to the horizontal plane of the middle surface of the cap in its reference configuration.

Key words. nonlinear shells, minimum problems, equilibrium stability

AMS(MOS) subject classifications. 34B15, 73H05, 73L99

1. Introduction. It is common for a thin and shallow elastic cap to be made to snap from a load-free and stress-free reference configuration into another equilibrium configuration, still load-free but no more stress-free, characterized by a macroscopic change in sign of the surface curvature.

A process through which such change in equilibrium configurations takes place is called an *eversion*, and the resulting configuration is called an *everted* equilibrium configuration. Indeed, the possibility of eversion of spherical shells, tubes, etc. is one of the mundane facts that best demonstrate the lack of uniqueness inherent in static problems of finite elasticity (cf. the discussion with illustrations given by Truesdell in [1]; the eversion into a spherical form of thick spherical shells has been considered by Antman in [2]).

Common experience shows how rather small disturbances induce the cap to snap back from the everted to the reference configuration; this suggests that the energy functional should attain a local minimum at the everted configuration.

Our paper is organized as follows. Section 2 is devoted to constructing the model of a spherical cap uniformly pulled along its rim by outward horizontal tractions. Section 3 covers the mathematical analysis of the resulting variational problem.

Care must be exercised in formulating this problem at a tractable level of generality, especially if we want to specify which simplifying hypotheses (among the many which shell theory practitioners use in a scattered and sometimes contradictory way) are going to be accepted, and under which form. Our formulation opens the way to tackle also other similar equilibrium problems, as it allows for a variety of meaningful loading and boundary conditions, in addition to those covered here. This is not the case for other related studies, where rather artificial boundary conditions appear to be essential for successful mathematical analysis.

A spherical cap has two aspect ratios, *thickness* and *shallowness*, defined by (2.1) and (2.2) below. In § 2.1, we give a succinct account of the kinematics of axially symmetric deformations of spherical caps of arbitrary thickness and shallowness. In

* Received by the editors October 16, 1987; accepted for publication (in revised form) August 1, 1988.
† Dipartimento di Ingegneria Civile Edile, Università di Roma 2, Via O. Raimondo, 00173 Roma, Italy.
‡ Istituto per le Applicazioni del Calcolo "M. Picone," Viale del Policlinico, 137, 00161 Roma, Italy.
§ Dipartimento di Matematica, Università di Roma 2, Via O. Raimondo, 00173 Roma, Italy.

§ 2.2, we carefully state the hypotheses underlying the notion of a *thin* cap. Roughly speaking, our thinness hypotheses, in the spirit of the semiinverse method, seek equilibrium displacements that have a particularly simple representation and that, in addition, give rise to strain states such that material fibers only suffer moderate stretching at the middle surface. In § 2.3 we stipulate that our cap is made of a homogeneous elastic material of the Saint-Venant & Kirchhoff type: this constitutive law, featuring a linear relationship between nonlinear measures of strain and stress, has the merit of simplicity (it also has drawbacks; cf. [3]). Integrating over the latitude and the thickness coordinates, we obtain the total energy functional as the sum of a stored energy functional and a loading potential; then, in §§ 2.4 and 2.5, respectively, we derive the relative field and boundary equations. Finally, in § 2.6, we give these equations the simplified form appropriate to *shallow* caps and arrive at Problem $\mathscr{P}$, i.e., a semilinear system of two ordinary differential equations of the second order, with linear boundary conditions, depending on two parameters: a parameter π_0, which is directly proportional to the applied load, and a parameter κ_0, which is inversely proportional to the cap's thickness.

In §§ 3.1–3.3 we select suitable weighted function spaces for Problem $\mathscr{P}$, introduce a pair of Green operators associated with the principal part operators in that problem, and end up reformulating it as a single functional equation whose solutions are the stationary points of a functional Γ depending parametrically on π_0 and κ_0. We then look for global (§ 3.3) and local (§ 3.4) minimizers of this functional. Our main findings may be summarized as follows.

For all values of π_0, that is, no matter whether the cap's rim is pulled or pushed, and for all possible values of the thickness parameter κ_0, there exists at least a global minimizer of Γ.

When the applied load is a sufficiently large outward traction, the global minimizer is unique (cf. Theorem 3(i)-(iv), and Theorem 4, § 3.3); moreover, for every $\kappa_0 > 0$ fixed, when $\pi_0 \to +\infty$ such global minimizer tends uniformly to a pseudoconfiguration of the cap that is not an equilibrium solution, but rather would correspond to "flattening" the cap into a disk under extremely high loads (cf. Thm. 5, § 3.3).

When the value of π_0 falls into the range of possible nonuniqueness for global minimizers and, in addition, κ_0 is greater than a positive number depending on π_0, the functional Γ is not convex (cf. Theorem 3(v)). In particular, Γ is not convex when $\pi_0 = 0$. In this situation, we find that for sufficiently thin caps there exists a sequence of local minimizers. This sequence, when $\kappa_0 \to +\infty$, tends to another pseudoconfiguration of the cap that is not an equilibrium solution, but may be described as the result of everting the reference configuration of an extremely thin cap with no bending resistance (in fact, a membrane) into its reflected image with respect to a horizontal support plane.

Thus, although we are unable to produce everted solutions explicitly, we believe we have sufficient grounds to claim that, at zero load, our model indeed accounts for eversion of sufficiently thin caps. This claim is substantiated by numerical computations presented in [4]; in particular, the everted shapes of Fig. 2 exhibit not only the overall change in curvature alluded to above, but also the expected flaring in the proximities of the cap's rim.

2. Formulation. Let $\{\mathbf{O}; \mathbf{i}, \mathbf{j}, \mathbf{k}\}$ be an orthonormal Cartesian frame, and let $\{R, \Theta, \Phi\}$ be a system of spherical coordinates centered at $\mathbf{O}$ (see Fig. 1).

Given three positive numbers R_0, ρ_0, and Θ_0, set

$$\varepsilon := \frac{\rho_0}{R_0} \tag{2.1}$$

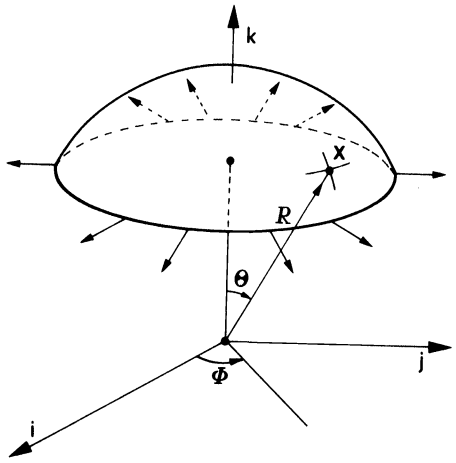


FIG. 1

and

$$\eta := \frac{1-\cos\Theta_0}{\sin\Theta_0}, \tag{2.2}$$

and consider the region $\mathscr{C}$

$$R \in]R_0-\rho_0, R_0+\rho_0[, \quad \Theta \in [0, \Theta_0[, \quad \Phi \in [0, 2\pi[, \tag{2.3}$$

of *thickness* 2ε and *shallowness* 2η. We will interpret $\mathscr{C}$ as the reference configuration of a spherical cap comprised of an elastic material of Saint-Venant & Kirchhoff type.

2.1. Kinematics. A displacement of $\mathscr{C}$ is a smooth vector field $\mathbf{u}$ over $\mathscr{C}$ such that the deformation

$$\mathbf{X} \mapsto \mathbf{x} = \mathbf{f}(\mathbf{X}) := \mathbf{X} + \mathbf{u}(\mathbf{X}) \tag{2.4}$$

is smooth, orientation preserving, and one to one;

$$\mathbf{F} := \nabla \mathbf{f} \quad \text{and} \quad \mathbf{H} := \nabla \mathbf{u} = \mathbf{F} - \mathbf{I}, \tag{2.5}$$

with $\mathbf{I}$ the gradient of the identity mapping, are the deformation and the displacement gradient, respectively. The linear and nonlinear deformation measures $\mathbf{E}$ and $\mathbf{D}$ are based on $\mathbf{H}$ and $\mathbf{F}$:

$$\mathbf{E} := \tfrac{1}{2}(\mathbf{H}+\mathbf{H}^T), \qquad \mathbf{D} := \tfrac{1}{2}(\mathbf{F}^T\mathbf{F}-\mathbf{I}) = \mathbf{E} + \tfrac{1}{2}\mathbf{H}^T\mathbf{H}. \tag{2.6}$$

Let $\mathbf{X} = X^1\mathbf{i} + X^2\mathbf{j} + X^3\mathbf{k}$ be a typical point of $\mathscr{C}$, having spherical coordinates

$$Z^1 = R, \quad Z^2 = \Theta, \quad Z^3 = \Phi. \tag{2.7}$$

At $\mathbf{X}$, the covariant basis is

$$\mathbf{E}_A := \partial_{Z^A}\mathbf{X}, \quad \text{for } A = 1, 2, 3; \tag{2.8}$$

as we will deal only with the so-called "physical" components of tensorial quantities throughout this paper, we introduce the associated orthonormal basis

$$\begin{gathered} \tilde{\mathbf{E}}_R := \mathbf{E}_R = \sin\Theta\mathbf{e} + \cos\Theta\mathbf{k}, \qquad \tilde{\mathbf{E}}_\Theta := R^{-1}\mathbf{E}_\Theta = \cos\Theta\mathbf{e} - \sin\Theta\mathbf{k}, \\ \tilde{\mathbf{E}}_\Phi := (R\sin\Theta)^{-1}\mathbf{E}_\Phi = \mathbf{e}^\perp, \end{gathered} \tag{2.9}$$

where the unit vectors $\mathbf{e}$ and $\mathbf{e}^\perp$ are defined to be

$$\mathbf{e} := \cos\Phi\mathbf{i} + \sin\Phi\mathbf{j}, \qquad \mathbf{e}^\perp := -\sin\Phi\mathbf{i} + \cos\Phi\mathbf{j}. \tag{2.10}$$

We here restrict attention to *axially symmetric* deformations. Accordingly, the displacement vector $\mathbf{u}$ has components

$$u_R := \mathbf{u} \cdot \tilde{\mathbf{E}}_R = u_R(R, \Theta), \qquad u_\Theta := \mathbf{u} \cdot \tilde{\mathbf{E}}_\Theta = u_\Theta(R, \Theta), \tag{2.11}$$
$$u_\Phi := \mathbf{u} \cdot \tilde{\mathbf{E}}_\Phi \equiv 0;$$

the displacement gradient $\mathbf{H}$ and the linear strain measure $\mathbf{E}$ have nonvanishing components

$$H_{RR} = u_{R,R}, \quad H_{\Theta\Theta} = \frac{1}{R}(u_{\Theta,\Theta} + u_R), \quad H_{\Phi\Phi} = \frac{1}{R}(u_R + \cot\Theta u_\Theta), \tag{2.12}$$
$$H_{R\Theta} = \frac{1}{R}(u_{R,\vartheta} - u_\Theta), \qquad H_{\Theta R} = u_{\Theta,R},$$

and

$$E_{RR} = H_{RR}, \quad E_{\Theta\Theta} = H_{\Theta\Theta}, \quad E_{\Phi\Phi} = H_{\Phi\Phi}, \tag{2.13}$$
$$E_{R\Theta} = \frac{1}{2}(H_{R\Theta} + H_{\Theta R}) = \frac{1}{2R}(u_{R,\Theta} + Ru_{\Theta,R} - u_\Theta),$$

respectively. The nonlinear strain measure $\mathbf{D}$ has nonvanishing components

$$D_{RR} = H_{RR} + \tfrac{1}{2}(H_{RR}^2 + H_{\Theta R}^2), \qquad D_{\Theta\Theta} = H_{\Theta\Theta} + \tfrac{1}{2}(H_{\Theta\Theta}^2 + H_{R\Theta}^2), \tag{2.14}$$
$$D_{\Phi\Phi} = H_{\Phi\Phi} + \tfrac{1}{2}H_{\Phi\Phi}^2, \qquad D_{R\Theta} = \tfrac{1}{2}(H_{R\Theta} + H_{\Theta R}) + \tfrac{1}{2}(H_{RR}H_{R\Theta} + H_{\Theta\Theta}H_{\Theta R}).$$

For axially symmetric deformations, it is easy to believe that $u_R(R, \cdot)$ and $u_\Theta(R, \cdot)$, provided their common domain of definition is extended in the obvious way, must be even and odd functions of Θ, respectively, for every fixed $R \in]R_0 - \rho_0, R_0 + \rho_0[$. Thus, in particular, all of $u_{R'}(R, \cdot)$, $u_\Theta(R, \cdot)$ and $H_{R\Theta}(R, \cdot)$ vanish at $\Theta = 0$.

2.2. Thin caps. To motivate some simplifying hypotheses that will prove crucial to our further developments, we now lay down counterparts adapted to the present context of the classical hypotheses of Kirchhoff's thin plate theory [5]-[9]. Our hypotheses are, for $R = R_0$,

$$D_{R\Theta} = D_{R\Phi} = 0, \tag{2.15}$$

$$D_{RR} = 0, \tag{2.16}$$

$$D_{\Theta\Theta} = D_{\Phi\Phi} = D_{\Theta\Phi} = 0. \tag{2.17}$$

As is well known, (2.15) and (2.16) express, respectively, the requirement that a radial material fiber remains, at the point where it crosses the middle surface, both orthogonal to the middle surface itself and unstretched. Equation (2.17) expresses the requirement that the material comprising the middle surface suffers no stretching in any conceivable deformation of the cap.

Note that for axially symmetric deformations, $(2.15)_2$ and $(2.17)_3$ are identically satisfied. In view of (2.14), the remaining equations can be written as

$$H_{R\Theta}^0(1 + H_{RR}^0) + H_{\Theta R}^0(1 + H_{\Theta\Theta}^0) = 0, \tag{2.18}$$

$$(1 + H_{RR}^0)^2 + (H_{\Theta R}^0)^2 = 1, \tag{2.19}$$

$$(1 + H_{\Theta\Theta}^0)^2 + (H_{R\Theta}^0)^2 = 1, \tag{2.20}$$

$$(1 + H_{\Phi\Phi}^0)^2 = 1. \tag{2.21}$$

(Here the superscript denotes evaluation at $R = R_0$.) This suggests that we look at situations when

$$|H^0_{RR}| \ll 1, \quad |H^0_{\Theta\Theta}| \ll 1, \quad |H^0_{\Phi\Phi}| \ll 1. \tag{2.22}$$

It follows from (2.22) that all of the above conditions of Kirchhoff's type are satisfied, except for (2.18), which reduces to

$$H^0_{R\Theta} + H^0_{\Theta R} \approx 0, \tag{2.23}$$

and the consistency condition

$$(H^0_{R\Theta})^2 \ll 1. \tag{2.24}$$

On adapting a procedure that has proved to be expedient in the case of plates ([8]), we look for solutions of (2.18)–(2.21) having the following form:

$$\begin{aligned} u_R(R, \Theta) &= u^0_R(\Theta) + \zeta(R) w(\Theta), \\ u_\Theta(R, \Theta) &= u^0_\Theta(\Theta) + \zeta(R) u(\Theta), \end{aligned} \qquad \zeta(R) := \frac{R - R_0}{R_0}, \tag{2.25}$$

implicitly restricting attention to *thin* caps.

We find that hypotheses (2.22) can be equivalently written as

$$\frac{1}{R_0}|w| \ll 1, \quad \frac{1}{R_0}|u^0_R + u^0_{\Theta,\Theta}| \ll 1, \quad \frac{1}{R_0}|u^0_R + \cot\Theta u^0_\Theta| \ll 1, \tag{2.26}$$

whereas (2.23) is satisfied if

$$u = u^0_\Theta - u^0_{R,\Theta}. \tag{2.27}$$

Accordingly, we write (2.25) as

$$u_R(R, \Theta) = u^0_R(\Theta), \qquad u_\Theta(R, \Theta) = u^0_\Theta(\Theta) + \zeta(R)(u^0_\Theta(\Theta) - u^0_{R,\Theta}(\Theta)). \tag{2.28}$$

Note that the parity properties requested for $u_R(R, \cdot)$ and $u_\Theta(R, \cdot)$ are guaranteed if we assume that

$$u^0_R(\cdot) \text{ is even and } u^0_\Theta(\cdot) \text{ is odd.} \tag{2.29}$$

From $(2.12)_{4,5}$ we obtain that

$$H_{\Theta R} = \frac{1}{R_0}(u^0_\Theta - u^0_{R,\Theta}) = H^0_{\Theta R} = -H^0_{R\Theta} = -H_{R\Theta}; \tag{2.30}_1$$

likewise, from $(2.12)_{1,2,3}$ it follows that

$$H_{RR} \approx 0, \tag{2.30}_2$$

$$H_{\Phi\Phi} \approx H^0_{\Phi\Phi} + \zeta \dot{H}^0_{\Phi\Phi} \quad \text{with } \dot{H}^0_{\Phi\Phi} = -\cot\Theta H^0_{R\Theta}, \tag{2.30}_3$$

$$H_{\Theta\Theta} \approx H^0_{\Theta\Theta} + \zeta \dot{H}^0_{\Theta\Theta} \quad \text{with } \dot{H}^0_{\Theta\Theta} = -H^0_{R\Theta,\Theta}. \tag{2.30}_4$$

We are now in a position to use (2.14) and (2.30) to evaluate the nonvanishing components of the nonlinear strain measure **D**. These are:

$$D_{\Theta\Theta} \approx D^0_{\Theta\Theta} + \zeta \dot{D}^0_{\Theta\Theta}, \qquad D_{\Phi\Phi} \approx D^0_{\Phi\Phi} + \zeta \dot{D}^0_{\Phi\Phi}, \tag{2.31}$$

with

$$\begin{aligned} D^0_{\Theta\Theta} &\approx H^0_{\Theta\Theta} + \frac{1}{2}(H^0_{R\Theta})^2 = \frac{1}{R_0}(u^0_R + u^0_{\Theta,\Theta}) + \frac{1}{2R_0^2}(u^0_\Theta - u^0_{R,\Theta})^2, \\ D^0_{\Phi\Phi} &\approx H^0_{\Phi\Phi} = \frac{1}{R_0}(u^0_R + \cot\Theta u^0_\Theta), \end{aligned} \tag{2.32}$$

and

$$\dot{D}^0_{\Theta\Theta} = \dot{H}^0_{\Theta\Theta} + (H^0_{\Theta\Theta} + H^0_{R\Theta})(\dot{H}^0_{\Theta\Theta} + \dot{H}^0_{R\Theta})$$

$$\approx \dot{H}^0_{\Theta\Theta} = \frac{1}{R_0}(u^0_\Theta - u^0_{R,\Theta})_{,\Theta}, \tag{2.33}$$

$$\dot{D}^0_{\Phi\Phi} = (1 + H^0_{\Phi\Phi})\dot{H}^0_{\Phi\Phi} \approx \dot{H}^0_{\Phi\Phi} = \frac{\cot\Theta}{R_0}(u^0_\Theta - u^0_{R,\Theta}).$$

2.3. Energy functional. For a Saint-Venant & Kirchhoff material, the stored energy density is

$$\sigma := \mu|\mathbf{D}|^2 + \tfrac{1}{2}\lambda(\operatorname{trace}(\mathbf{D}))^2, \tag{2.34}$$

with μ and λ two constant material moduli;

$$\Sigma := \int_{\mathscr{C}} \sigma \tag{2.35}$$

is the associated stored energy functional. The following constitutive law:

$$\mathbf{S} := \partial_{\mathbf{D}}\sigma = 2\mu\mathbf{D} + \lambda \operatorname{trace}(\mathbf{D})I \tag{2.36}$$

delivers the stress accompanying $\mathbf{D}$.

In view of (2.31)-(2.33), the integrand in (2.35) can be written as follows:

$$\sigma(R, \Theta) \approx \sigma^0(\Theta) + \zeta(R)\dot{\sigma}^0(\Theta) + \tfrac{1}{2}\zeta^2(R)\ddot{\sigma}^0(\Theta), \tag{2.37}$$

where

$$\begin{aligned}
\sigma^0 &= \tfrac{1}{2}(2\mu+\lambda)((D^0_{\Theta\Theta})^2 + (D^0_{\Phi\Phi})^2) + \lambda D^0_{\Theta\Theta}D^0_{\Phi\Phi},\\
\dot{\sigma}^0 &= (2\mu+\lambda)(D^0_{\Theta\Theta}\dot{D}^0_{\Theta\Theta} + D^0_{\Phi\Phi}\dot{D}^0_{\Phi\Phi}) + \lambda(D^0_{\Theta\Theta}\dot{D}^0_{\Phi\Phi} + D^0_{\Phi\Phi}\dot{D}^0_{\Theta\Theta}),\\
\ddot{\sigma}^0 &= (2\mu+\lambda)((\dot{D}^0_{\Theta\Theta})^2 + (\dot{D}^0_{\Phi\Phi})^2) + 2\lambda\dot{D}^0_{\Theta\Theta}\dot{D}^0_{\Phi\Phi}.
\end{aligned} \tag{2.38}$$

Integrating over the latitude and the thickness, we obtain

$$\begin{aligned}
\Sigma = \frac{4\pi}{3}R_0^3\Bigg(&\varepsilon(3+\varepsilon^2)\int_0^{\Theta_0}\sigma^0\sin\Theta\,d\Theta + 2\varepsilon^3\int_0^{\Theta_0}\dot{\sigma}^0\sin\Theta\,d\Theta\\
&+\frac{1}{2}\varepsilon^3\left(1+\frac{3}{5}\varepsilon^2\right)\int_0^{\Theta_0}\ddot{\sigma}^0\sin\Theta\,d\Theta\Bigg),
\end{aligned} \tag{2.39}$$

where ε is the thickness parameter (2.1). Adhering to a common practice (cf., e.g., [10]), we take

$$\Sigma \approx \frac{4\pi}{3}R_0^3\left(3\varepsilon\int_0^{\Theta_0}\sigma^0\sin\Theta\,d\Theta + \frac{1}{2}\varepsilon^3\int_0^{\Theta_0}\ddot{\sigma}^0\sin\Theta\,d\Theta\right). \tag{2.40}$$

We stipulate that the cap is pulled along its rim $\mathscr{R} := \{\mathbf{X}\in\partial\mathscr{C}\,|\,\Theta = \Theta_0\}$ by a horizontal outward traction $\mathbf{p}$, of constant magnitude p per unit area,

$$\mathbf{p} = p\mathbf{e}, \tag{2.41}$$

with $\mathbf{e}$ defined by $(2.10)_1$. The associated loading potential is

$$\Pi := \int_{\mathscr{R}} \mathbf{p}\cdot\mathbf{u}; \tag{2.42}$$

by use of (2.11) and (2.28), integrating again over the latitude and the thickness, we arrive at

$$\Pi = p\frac{4\pi}{3} R_0^2 \sin\Theta_0 (3\varepsilon(\sin\Theta_0 u_R^0(\Theta_0) + \cos\Theta_0 u_\Theta^0(\Theta_0)) + \varepsilon^3 \cos\Theta_0 (u_\Theta^0(\Theta_0) - u_{R,\Theta}^0(\Theta_0))). \tag{2.43}$$

Suppose now that we would be willing to stipulate that the vertical component of the displacement vector vanish along the cap's rim:

$$\mathbf{u}\cdot\mathbf{k} = 0 \quad \text{for all points of } \mathscr{R}. \tag{2.44}$$

This assumption would imply that

$$\cos\Theta_0 u_R^0(\Theta_0) - \sin\Theta_0 u_\Theta^0(\Theta_0) = 0, \qquad \sin\Theta_0 (u_\Theta^0(\Theta_0) - u_{R,\Theta}^0(\Theta_0)) = 0. \tag{2.45}$$

On recalling $(2.30)_1$ we note that, as $\sin\Theta_0 \neq 0$, the geometric condition $(2.45)_2$ is equivalent to

$$H_{\Theta R}^0(\Theta_0) = 0. \tag{2.46}$$

Motivated by the above argument, we choose to reduce the loading potential (2.43) to

$$\Pi = p 4\pi R_0^2 \varepsilon \sin\Theta_0 (\sin\Theta_0 u_R^0(\Theta_0) + \cos\Theta_0 u_\Theta^0(\Theta_0)), \tag{2.47}$$

stipulate that $(2.45)_1$ holds, and dispense with $(2.45)_2$.

In conclusion, we will request that the energy functional

$$I\{u_R^0, u_\Theta^0\} := \Sigma\{u_R^0, u_\Theta^0\} - \Pi\{u_R^0, u_\Theta^0\}, \tag{2.48}$$

with Σ and Π given by (2.40) and (2.47), respectively, be stationary over the class of variations (v_R, v_Θ) obeying the geometric condition $(2.45)_1$:

$$\cos\Theta_0 v_R(\Theta_0) - \sin\Theta_0 v_\Theta(\Theta_0) = 0. \tag{2.49}$$

Condition (2.49) guarantees that the variation vector

$$\mathbf{v} := v_R \mathbf{E}_R + v_\Theta \mathbf{E}_\Theta$$

has null vertical component at the rim. We will also stipulate that $v_R(\cdot)$, like $u_R(R, \cdot)$, is an even function, and that $v_\Theta(\cdot)$, like $u_\Theta(R, \cdot)$ is an odd function of Θ.

2.4. Euler–Lagrange equations. After some manipulations the following system of differential equations in $]0, \Theta_0[$ is arrived at:

$$\begin{aligned} &(\sin\Theta M' + \cos\Theta(M-N) - \sin\Theta H_{\Theta R}^0 S)' - \sin\Theta(S+T) = 0, \\ &\sin\Theta M' + \cos\Theta(M-N) - \sin\Theta H_{\Theta R}^0 S - \cos\Theta T + (\sin\Theta S)' = 0, \end{aligned} \tag{2.50}$$

where for convenience we have denoted differentiation with respect to Θ by an apostrophe and we have set

$$\begin{aligned} &S := D_{\Theta\Theta}^0 + \frac{\lambda}{2\mu}(D_{\Theta\Theta}^0 + D_{\Phi\Phi}^0), \qquad T := D_{\Phi\Phi}^0 + \frac{\lambda}{2\mu}(D_{\Theta\Theta}^0 + D_{\Phi\Phi}^0), \\ &\frac{3}{\varepsilon^2} M := \dot{D}_{\Theta\Theta}^0 + \frac{\lambda}{2\mu}(\dot{D}_{\Theta\Theta}^0 + \dot{D}_{\Phi\Phi}^0), \qquad \frac{3}{\varepsilon^2} N := \dot{D}_{\Phi\Phi}^0 + \frac{\lambda}{2\mu}(\dot{D}_{\Theta\Theta}^0 + \dot{D}_{\Phi\Phi}^0). \end{aligned} \tag{2.51}$$

To give (2.50) a more tractable form we need some preliminary results. First, from (2.32) and $(2.51)_{1,2}$ we obtain that

(2.52)
$$\sin\Theta T'+\cos\Theta(T-S)-\frac{\lambda}{2(\lambda+\mu)}\sin\Theta(T+S)'=-\cos\Theta H^0_{\Theta R}\left(\frac{1}{2}H^0_{\Theta R}+\tan\Theta\right).$$

Second, in view of (2.33) and $(2.51)_{3,4}$, we have that

(2.53)
$$\sin\Theta M'+\cos\Theta(M-N)=\frac{\varepsilon^2}{3}\sin\Theta\Big(H^{0\prime\prime}_{\Theta R}+\cot\Theta(H^{0\prime}_{\Theta R}-\cot\Theta H^0_{\Theta R})+\frac{\lambda}{2\mu}(H^{0\prime}_{\Theta R}+\cot\Theta H^0_{\Theta R})'\Big).$$

Third, solving $(2.50)_2$ for T, we get

$$T=S+\tan\Theta S'+\frac{1}{\cos\Theta}(\sin\Theta M'+\cos\Theta(M-N)-\sin\Theta H^0_{\Theta R}S), \tag{2.54}$$

and substituting (2.54) into $(2.50)_1$ yields

$$(\sin^2\Theta S)'=(\cos\Theta(\sin\Theta M'-\cos\Theta(M-N)-\sin\Theta H^0_{\Theta R}S))'. \tag{2.55}$$

Integrating (2.55) over $[0,\Theta]$, we obtain

$$\tan^2\Theta S=\frac{1}{\cos\Theta}(\sin\Theta M'+\cos\Theta(M-N)-\sin\Theta H^0_{\Theta R}S) \tag{2.56}$$

(here we have made use of (2.61) below); but, (2.54) and (2.56) together imply that

$$T=(\tan\Theta S)'. \tag{2.57}$$

Now, from (2.52) and (2.57), we deduce

$(2.58)_1$
$$-\tan\Theta\left(\frac{\lambda+2\mu}{2(\lambda+\mu)}(\tan\Theta S)''-\frac{\lambda}{2(\lambda+\mu)}S'\right)+(S-(\tan\Theta S)')=H^0_{\Theta R}\left(\frac{1}{2}H^0_{\Theta R}+\tan\Theta\right);$$

from $(2.50)_2$, (2.53), and (2.57), we have

$(2.58)_2$
$$\frac{\varepsilon^2}{3}\left(H^{0\prime\prime}_{\Theta R}+\cot\Theta(H^{0\prime}_{\Theta R}-\cot\Theta H^0_{\Theta R})+\frac{\lambda}{2\mu}(H^{0\prime}_{\Theta R}+\cot\Theta H^0_{\Theta R})'\right)=(H^0_{\Theta R}+\tan\Theta)S.$$

The ordinary second-order equations $(2.58)_{1,2}$ compose a system for the unknowns S and $H^0_{\Theta R}$, expressing the equilibrium of the cap under study.

2.5. Boundary conditions. In order that the first variation of the energy functional (2.48) vanish identically, we must require that the following condition be satisfied at the boundary:

$$[\sin\Theta S v_\Theta-\sin\Theta H^0_{\Theta R}Sv_R+\sin\Theta M(v_\Theta-v_{R'})+(\sin\Theta M'+\cos\Theta(M-N))v_R]_0^{\Theta_0}$$
$$-\tilde{\pi}_0\sin\Theta_0(\sin\Theta_0 v_R(\Theta_0)+\cos\Theta_0 v_\Theta(\Theta_0))=0 \tag{2.59}$$

for all admissible variations v_R and v_Θ, and for

$$\tilde{\pi}_0:=\frac{p}{2\mu},\qquad \mu>0. \tag{2.60}$$

On evaluating at $\Theta = 0$ the term between square brackets in (2.59), we see that it vanishes if and only if

$$M(0) - N(0) = 0. \tag{2.61}$$

As

$$M(\Theta) - N(\Theta) = \dot{D}^0_{\Theta\Theta}(\Theta) - \dot{D}^0_{\Phi\Phi}(\Theta) = -\frac{(\sin\Theta H^0_{\Theta R})'}{\sin\Theta},$$

the parity properties of axially symmetric solutions imply that indeed

$$\lim_{\Theta\to 0}(M(\Theta) - N(\Theta)) = 0.$$

Taking (2.49) and (2.61) into account, we have that (2.59) yields the two natural boundary conditions that prevail at the rim; these are

$$M(\Theta_0) = 0 \quad \text{and} \quad \sin\Theta_0 Q(\Theta_0) + \cos\Theta_0 S(\Theta_0) - \tilde{\pi}_0 = 0, \tag{2.62}$$

and involve the "bending moment" M and the "shear force" Q, the latter being defined by

$$\sin\Theta Q := \sin\Theta M' + \cos\Theta(M - N) - \sin\Theta H^0_{\Theta R} S. \tag{2.63}$$

In view of (2.33) and $(2.51)_3$, $(2.62)_1$ can be written as follows:

$$H^{0\prime}_{\Theta R}(\Theta_0) + \frac{\lambda}{\lambda + 2\mu}\cot\Theta_0 H^0_{\Theta R}(\Theta_0) = 0; \tag{2.64}_1$$

likewise, as (2.56) and (2.63) imply that

$$Q = \tan\Theta S \tag{2.65}$$

at equilibrium, $(2.62)_2$ can be written as follows:

$$S(\Theta_0) - \tilde{\pi}_0\cos\Theta_0 = 0. \tag{2.64}_2$$

The boundary conditions prevailing at $\Theta = 0$ are dictated by the often-recalled parity properties of axially symmetric deformations, as they are reflected in such constructs as S and $H^0_{\Theta R}$. Indeed, it is easy to see from the relevant definitions that S and $H^0_{\Theta R}$ must be even and odd functions of Θ, respectively. Therefore, we insist that

$$S'(0) = 0 \quad \text{and} \quad H^0_{\Theta R}(0) = 0. \tag{2.66}$$

2.6. Shallow caps. The differential equations (2.58), supplemented by the boundary conditions (2.64) and (2.66), regulate the equilibrium of a thin spherical cap uniformly pulled at its rim. The corresponding equations for the case of *shallow* caps, i.e., when $2\eta \approx \Theta_0$, are obtained by simply replacing throughout $\tan\Theta$ with Θ, etc. The resulting differential system is

$$\begin{aligned} F'' + \frac{3}{\Theta}F' &= \kappa_0(1+F)G + 2\pi_0\kappa_0^2(1+F), \\ G'' + \frac{3}{\Theta}G' &= -\kappa_0(2+F)F, \end{aligned} \tag{2.67}$$

where for convenience we have set

$$\begin{aligned} &\Theta F := H^0_{\Theta R}, \qquad G := \frac{2\kappa_0}{1+\nu}(S - \tilde{\pi}_0), \\ &2\kappa_0^2 := \frac{3}{\varepsilon^2}(1-\nu^2), \qquad \pi_0 := \frac{\tilde{\pi}_0}{1+\nu} \quad \text{with } \nu := \frac{\lambda}{\lambda+2\mu}. \end{aligned} \tag{2.68}$$

The boundary conditions are

$$F'(0)=0, \qquad \Theta_0 F'(\Theta_0)+(1+\nu)F(\Theta_0)=0,$$
$$G'(0)=0, \qquad G(\Theta_0)=0, \tag{2.69}$$

where $(2.69)_1$ efficiently replaces $(2.66)_2$, a condition which is rendered empty by the change of variable $(2.68)_1$.

3. Analysis. We will refer to the boundary value problem (2.67), (2.69) as to Problem $\mathcal{P}$. We remark that ν is the only material parameter of importance in this problem. For Saint-Venant & Kirchhoff materials, ν is the direct counterpart of Poisson's modulus for isotropic, linearly elastic materials. For reasons of physical plausibility we assume that

$$\nu \in]-1, \tfrac{1}{2}[; \tag{3.1}$$

hence, definitions $(2.68)_{2,3,4}$ make sense, and we may regard the thickness parameter in Problem $\mathcal{P}$ as positive:

$$\kappa_0 > 0. \tag{3.2}$$

On the other hand, the sign of π_0, the applied force parameter, tells us whether the cap is pulled ($\pi_0 > 0$) or pushed ($\pi_0 < 0$) at its rim.

In this section we will study existence, multiplicity, and stability of solutions to Problem $\mathcal{P}$, for ν fixed, and for the parameters κ_0, π_0 varying over their admissible ranges. We will devote special attention to the case of zero load ($\pi_0 = 0$), when eversion is in order.

3.1. Function spaces. To motivate our choice of function spaces, consider first, in an informal way, the linear differential operator that composes the principal part of both equations (2.67):

$$-L[\,\cdot\,] := [\,\cdot\,]'' + 3\Theta^{-1}[\,\cdot\,]' = \Theta^{-3}[\Theta^3[\,\cdot\,]']'. \tag{3.3}$$

For h some "test" function vanishing in a neighborhood of zero, the weak version of the equation

$$L[f] = g \tag{3.4}$$

is

$$-[\Theta^3 f' h]_0^{\Theta_0} + \int_0^{\Theta_0} \Theta^3 f' h' \, d\Theta = \int_0^{\Theta_0} \Theta^3 g h \, d\Theta, \tag{3.5}$$

or rather, taking the boundary condition $(2.69)_2$ into account,

$$(1+\nu)\Theta_0^2 f(\Theta_0) h(\Theta_0) + \int_0^{\Theta_0} \Theta^3 f' h' \, d\Theta = \int_0^{\Theta_0} \Theta^3 g h \, d\Theta. \tag{3.6}$$

In the light of the above, we introduce the following weighted Hilbert spaces:

K = the space of all functions on $]0, \Theta_0[$ that are square integrable with respect to the weight Θ^3,

K^1 = the space of all functions on $]0, \Theta_0]$ whose first derivatives are elements of K,

K_0^1 = the space of all functions of K^1 vanishing at Θ_0,

with scalar product and norm

$$(f_1, f_2)_0 := \int_0^{\Theta_0} \Theta^3 f_1(\Theta) f_2(\Theta) \, d\Theta, \quad \|f\|_0^2 := (f,f)_0 \quad \text{in } K, \tag{3.7}$$

$$(f_1, f_2)_1 := (f_1', f_2')_0 + (1+\nu)\Theta_0^2 f_1(\Theta_0) f_2(\Theta_0), \qquad \|f\|_1^2 := (f,f)_1 \text{ in } K^1. \tag{3.8}$$

We point out that K_0^1 turns out to be the orthogonal complement in K^1 of the subspace of all constants; thus, our choice of function spaces nicely accommodates the boundary conditions $(2.69)_{3,4}$.

LEMMA 1. *Let $f \in K^1$. Then, there exist two positive constants γ_1 and γ_2 such that*

(i) $\int_0^{\Theta_0} \Theta f^2(\Theta)\, d\Theta \leqq \gamma_1 \|f\|_1^2$;

(ii) $\int_0^{\Theta_0} \Theta^3 f^4(\Theta)\, d\Theta \leqq \gamma_2 \|f\|_1^4$.

Moreover,

(iii) *K^1 is compactly embedded into K.*

Proof. Preliminarily, note that both (i) and (ii) trivially hold true if f has constant value. Hence, it is sufficient to consider the case when $f \in K_0^1$. In this case, we have that

$$f^2(\Theta) = -2 \int_\Theta^{\Theta_0} f(\Phi) f'(\Phi)\, d\Phi,$$

and

$$f^4(\Theta) \leqq 4 \left(\int_\Theta^{\Theta_0} f^2(\Phi)\, d\Phi \right) \left(\int_\Theta^{\Theta_0} f'^2(\Phi)\, d\Phi \right).$$

Consequently,

$$\Theta f^2(\Theta) \leqq 2 \int_\Theta^{\Theta_0} \Phi |f(\Phi) f'(\Phi)|\, d\Phi \tag{3.9}_1$$

and

$$\Theta^3 f^4(\Theta) \leqq 4 \left(\int_\Theta^{\Theta_0} f^2(\Phi)\, d\Phi \right) \left(\int_\Theta^{\Theta_0} \Phi^3 f'^2(\Phi)\, d\Phi \right) \leqq 4 \|f\|_1^2 \int_\Theta^{\Theta_0} f^2(\Phi)\, d\Phi. \tag{3.9}_2$$

Consider now the identity

$$\int_0^{\Theta_0} \Theta f(\Theta)\, d\Theta = \int_0^{\Theta_0} d\Theta \int_\Theta^{\Theta_0} f(\Phi)\, d\Phi, \tag{3.10}$$

that holds whenever the mapping $\Theta \mapsto \Theta f(\Theta)$ is integrable over $]0, \Theta_0[$. In view of (3.10), $(3.9)_1$ yields

$$\begin{aligned} \int_0^{\Theta_0} \Theta f^2(\Theta)\, d\Theta &\leqq 2 \int_0^{\Theta_0} \Theta^2 |f(\Theta) f'(\Theta)|\, d\Theta \\ &\leqq 2 \left(\int_0^{\Theta_0} \Theta f^2(\Theta)\, d\Theta \right)^{1/2} \left(\int_0^{\Theta_0} \Theta^3 f'^2(\Theta)\, d\Theta \right)^{1/2} \end{aligned}$$

which is (i). Item (ii) is proved in a completely analogous way, using the last formula above and $(3.9)_2$.

To establish (iii), consider a bounded sequence $\{f_n\} \subset K^1$. As all functions f_n are equicontinuous over every subinterval $[\varepsilon, \Theta_0]$ with $\varepsilon > 0$, we can always find a subsequence converging uniformly in any one of those subintervals. The desired result then follows from the estimate under (i). □

3.2. Green operators. As a consequence of Lemma 1(i) $f \mapsto f^2$ is a continuous mapping from K^1 into K. Therefore, in Problem $\mathcal{P}$, the right-hand sides of equations (2.67) belong to K if both G and F belong to K^1.

Then pick $g \in K$, and consider the following two problems associated with (equation (3.4) and) Problem $\mathscr{P}$:

($\mathscr{P}_0$) Find $f \in K_0^1$ such that $(f, h)_1 = (g, h)_0$ for all $h \in K_0^1$.

($\mathscr{P}_1$) Find $f \in K^1$ such that $(f, h)_1 = (g, h)_0$ for all $h \in K^1$.

On appealing to standard results from the Hilbertian theory for boundary value problems, we see that Problem $\mathscr{P}_\alpha$ ($\alpha = 0, 1$) has a unique solution f_α that solves (3.4); for almost every $\Theta \in]0, \Theta_0[$, satisfies the boundary condition in Θ_0 (i.e., either $f_0(\Theta_0) = 0$ for $\alpha = 0$ or $\Theta_0 f_1'(\Theta_0) + (1+\nu) f_1(\Theta_0) = 0$ for $\alpha = 1$), and satisfies the boundary condition in 0 in a weak sense (i.e., $f_\alpha' \in K$). Accordingly, the mappings $g \mapsto f_\alpha$ ($\alpha = 0, 1$) define two (linear and bounded) Green operators G_α from K into, respectively, K_0^1 for $\alpha = 0$ and K^1 for $\alpha = 1$; we have also that, for all functions $h \in K_0^1$ for $\alpha = 0$ and all functions $h \in K^1$ for $\alpha = 1$,

$$(G_\alpha[g], h)_1 = (g, h)_0. \tag{3.11}$$

LEMMA 2. *Let G_α be either one of the Green operators defined just above, and let g, f_α ($\alpha = 0, 1$) be such that $f_\alpha = G_\alpha[g]$. Then, for $\Theta \in]0, \Theta_0]$, the following inequalities hold true*:

(i) $|f_\alpha'(\Theta)|^2 \leqq \frac{1}{6} \int_0^\Theta \Theta |g(\Theta)|^2 \, d\Theta$;

(ii) $|f_\alpha(\Theta_0) - f_\alpha(\Theta)| \leqq \frac{1}{2} \|g\|_0 \ln(\Theta_0/\Theta)$.

Moreover,

(iii) *G_α maps compactly K into K^1;*

(iv) *G_α can be seen as a self-adjoint operator of K into itself;*

(v) *G_α is a positive operator, i.e., $g > 0 \Rightarrow G_\alpha[g] = f_\alpha > 0$;*

(vi) *G_α maps continuously $C^0[0, \Theta_0]$ into $C^2[0, \Theta_0]$.*

Proof. Preliminarily, we observe that it follows from the strong integrability requirement necessary for the right-hand side of (i) to be finite that the boundary condition at zero is satisfied in classical sense, i.e.,

$$\lim_{\Theta \to 0+} f_\alpha'(\Theta) = 0.$$

Under the current hypotheses,

$$(G_\alpha[g](\Theta))' = f_\alpha'(\Theta) = -\Theta^{-3} \int_0^\Theta \Phi^3 g(\Phi) \, d\Phi. \tag{3.12}$$

To see this, first take $g(\Theta) \equiv 0$ in an arbitrarily small right neighborhood of zero; as $f_\alpha(\Theta)$ has constant value in such a neighborhood, (3.12) holds; a straightforward density argument then shows that this conclusion continues to hold for an arbitrary $g \in K$.

By integrating $(3.12)_2$ over $[\Theta, \Theta_0]$, we obtain

$$f_\alpha(\Theta) = f_\alpha(\Theta_0) - \frac{1}{2}\Theta_0^{-2} \int_0^{\Theta_0} \Phi^3 g(\Phi) \, d\Phi + \frac{1}{2}\Theta^{-2} \int_0^\Theta \Phi^3 g(\Phi) \, d\Phi + \frac{1}{2} \int_\Theta^{\Theta_0} \Phi g(\Phi) \, d\Phi. \tag{3.13}$$

Now, for $\alpha = 0$, $f_0(\Theta_0) = 0$ and (3.13) yields

$$f_0(\Theta) = \frac{1}{2} \int_0^\Theta \Phi^3 (\Theta^{-2} - \Theta_0^{-2}) g(\Phi) \, d\Phi + \frac{1}{2} \int_\Theta^{\Theta_0} \Phi^3 (\Phi^{-2} - \Theta_0^{-2}) g(\Phi) \, d\Phi; \tag{3.14}$$

for $\alpha = 1$, again $(3.12)_2$ implies that

$$\int_0^{\Theta_0} \Phi^3 g(\Phi)\, d\Phi = -\Theta_0^3 f_1'(\Theta_0) = \Theta_0^2(1+\nu) f_1(\Theta_0), \tag{3.15}$$

so that (3.13) this time yields

$$f_1(\Theta) = f_0(\Theta) + (1+\nu)^{-1}\Theta_0^{-2} \int_0^{\Theta_0} \Phi^3 g(\Phi)\, d\Phi. \tag{3.16}$$

From $(3.12)_2$, (i) follows by splitting $\Theta^3 g(\Theta)$ as $\Theta^{5/2} \times \Theta^{1/2} g(\Theta)$ and using the Schwartz inequality. Next, by integrating $(3.12)_2$ over $[\Theta, \Theta_0]$ and using the Schwartz inequality again, we get (ii).

The proofs of statements (iii) and (iv) are almost trivial, and we omit them. The proof of statement (v) easily follows from either a glance to (3.14) and (3.16) or a direct application of the Maximum Principle.

Finally, assume that g is continuous at zero. To show that statement (vi) holds true, it is sufficient to divide both sides of $(3.12)_2$ by Θ, and then pass to the limit for $\Theta \to 0$, to get

$$f_\alpha''(0) = \tfrac{1}{4} g(0). \qquad \square$$

We are now in a position to write Problem $\mathscr{P}$ in the following form:

$$(\mathscr{P}') \qquad \begin{aligned} F &= -\kappa_0 G_1[(1+F)G + 2\pi_0\kappa_0(1+F)], \\ G &= \kappa_0 G_0[(2+F)F]. \end{aligned}$$

We remark that the inequalities under Lemma 2(i) and (ii) ensure us that every solution of Problem $\mathscr{P}'$ is in fact a classical solution of Problem $\mathscr{P}$.

3.3. Global minimizers. We begin by noting that Problem $\mathscr{P}'$ can be written under the form of a single equation for the unknown F, namely,

$$F + \kappa_0^2 G_1[(1+F)G_0[(2+F)F] + 2\pi_0(1+F)] = 0. \tag{3.17}$$

On recalling (3.11), $(3.12)_1$, and Lemma 2(iv), it can be shown that the left-hand side of (3.17) is the Fréchet derivative of the functional

$$\Gamma(F) := \frac{1}{2}\|F\|_1^2 + \frac{\kappa_0^2}{4}\|G_0[(2+F)F]\|_1^2 + \pi_0\kappa_0^2(2+F, F)_0; \tag{3.18}$$

consequently, the solutions of (3.17) are the stationary points of Γ. Moreover, the second differential of Γ at F turns out to be the quadratic form

$$\Gamma''(F; h) := \|h\|_1^2 + 2\kappa_0^2\|G_0[(1+F)h]\|_1^2 + \kappa_0^2(G_0[(2+F)F], h^2)_0 + 2\pi_0\kappa_0^2\|h\|_0^2. \tag{3.19}$$

THEOREM 3. *The functional* Γ *over* K^1 *defined by* (3.18) *has the following properties*:

(i) Γ *is lower semicontinuous with respect to the weak topology of* K^1;

(ii) Γ *is coercive, i.e.,* $\Gamma(F) \to +\infty$ *as* $\|F\|_1 \to +\infty$;

(iii) Γ *is strictly convex for* $\pi_0 \geqq \Theta_0^2/16$;

(iv) *for every* $\pi_0 < \Theta_0^2/16$ *there exists a positive number* κ_{π_0} *such that* Γ *is not convex for* $\kappa_0 > \kappa_{\pi_0}$.

Proof. Statement (i) is an easy consequence of (iii) in Lemma 2.

Statement (ii) is proved by contradiction. Indeed, let $\{f_n\} \subset K^1$, with $\|f_n\| \to +\infty$, and assume that there exists some constant γ such that

$$\Gamma(f_n) \leqq \gamma. \tag{3.20}$$

From (3.18) we then have that

$$\lim_{n\to\infty} \|f_n\|^{-4}\Gamma(f_n) = \lim_{n\to\infty} \left\| G_0\left[\frac{f_n^2}{\|f_n\|_1^2}\right]\right\|_1^2 = 0. \tag{3.21}$$

As zero is not a proper value of G_0, it follows from $(3.21)_2$ that

$$\frac{\|f_n^2\|_0}{\|f_n\|_1^2} \to 0.$$

Consequently,

$$\liminf_{n\to\infty} \Gamma(f_n) \geqq \tfrac{1}{2} \liminf_{n\to\infty} \|f_n\|_1^2 = +\infty,$$

an inequality incompatible with (3.20).

To prove items (iii) and (iv), we begin by observing that

$$\Gamma''(F; h) - \Gamma''(-1; h) = 2\kappa_0^2 \|G_0[(1+F)h]\|_1^2 + \kappa_0^2 (G_0[(1+F)^2], h^2)_0 \geqq 0, \tag{3.22}$$

where the last inequality follows from Lemma 2(v). In addition, we observe that

$$\Gamma''(-1; h) = \|h\|_1^2 + 2\pi_0\kappa_0^2 \|h\|_0^2 - \kappa_0^2 (G_0[1], h^2)_0, \tag{3.23}$$

and that

$$G_0[1] = \frac{\Theta_0^2 - \Theta^2}{8}. \tag{3.24}$$

But, (3.22)-(3.24) together imply that

$$\Gamma''(F; h) \geqq \|h\|_1^2 + \kappa_0^2 \int_0^{\Theta_0} \Theta^3 \left(2\pi_0 - \frac{\Theta_0^2 - \Theta^2}{8}\right) h^2(\Theta)\, d\Theta; \tag{3.25}$$

from (3.25), the desired conclusions easily follow. □

As a straightforward corollary of Theorem 3 we have the following theorem.

THEOREM 4. *For all admissible values of the parameters, i.e., for all* $\pi_0 \in \mathbb{R}$ *and* $\kappa_0 > 0$, *a global minimizer exists in* K^1 *for the functional* Γ; *moreover, if* $\pi_0 \geqq \Theta_0^2/16$, *such a minimizer is unique.*

Thus, the equilibrium problem under study has at least a solution of minimum energy, which is the only solution when sufficiently large outward tractions are applied along the cap's rim.

Our next theorem gives qualitative information concerning the nature of global minimizers.

THEOREM 5. *Let* $\kappa_0 > 0$ *be fixed*; *moreover, for every choice of* $\pi_0 \geqq \Theta_0^2/16$, *let* F_{π_0} *denote the global minimizer of* Γ *associated with* (κ_0 *and*) π_0. *Then, as* $\pi_0 \to +\infty$, $(F_{\pi_0}(\Theta) + 1) \to 0$ *uniformly in* $[0, \Theta_0]$.

Proof. Set $H_{\pi_0} := F_{\pi_0} + 1$. Then, from the second equation in Problem $\mathscr{P}'$, we get that

$$G_{\pi_0} = \kappa_0 G_0[H_{\pi_0}^2] - \kappa_0 G_0[1]; \tag{3.26}$$

in turn, in view of Lemma 2(v), (3.24), and the present hypothesis on the parameter π_0, (3.26) implies that

$$G_{\pi_0} + 2\pi_0\kappa_0 > 0. \tag{3.27}$$

With this, from $(2.67)_1$ we obtain

$$-M[F_{\pi_0}] := -(L + \kappa_0(G_{\pi_0} + 2\pi_0\kappa_0)I)[F_{\pi_0}] = \kappa_0(G_{\pi_0} + 2\pi_0\kappa_0) \geqq 0, \tag{3.28}$$

where we have denoted by I the identity operator. Now, by the Maximum Principle,

$$F_{\pi_0}(\Theta) \leqq 0. \tag{3.29}$$

On the other hand, as

$$L[H_{\pi_0}] \geqq 0,$$

and as

$$H'_{\pi_0}(0) = 0, \qquad \Theta_0 H'_{\pi_0}(\Theta_0) + (1+\tilde{\nu}) H_{\pi_0}(\Theta_0) > 0,$$

by the Maximum Principle the global minimizer F_{π_0} must be such that $F_{\pi_0}(\Theta) \geqq -1$. We then conclude, with the aid of (3.29), that

$$-1 \leqq F_{\pi_0}(\Theta) \leqq 0. \tag{3.30}$$

Suppose now that there exists a sequence $\{\pi_{0,n}\} \to +\infty$ such that

$$\max_{[0,\Theta_0]} F_{\pi_{0,n}}(\Theta) \geqq \alpha > -1. \tag{3.31}$$

We may safely assume, in addition, that

$$\pi_{0,n+1} - \pi_{0,n} \geqq \Theta_0^2/16.$$

Therefore, repeating the argument leading to (3.29), we have that the sequence $\{F_{\pi_{0,n}}\}$ decreases pointwise to some function F_∞; we also have that, in the topologies of K and K^1, respectively,

$$\{(2+F_{\pi_{0,n}})F_{\pi_{0,n}}\} \to (2+F_\infty)F_\infty, \qquad \{G_{\pi_{0,n}}\} \to \kappa_0 G_0[(2+F_\infty)F_\infty].$$

In view of this, dividing the first equation in Problem $\mathscr{P}'$ by $\pi_{0,n}$ and taking the limit for $n \to \infty$, it follows that $G_1[F_\infty + 1] = 0$, which in turn implies that $F_\infty + 1 = 0$. But then, by Dini's Theorem, $F_{\pi_{0,n}}(\Theta) \to -1$ uniformly; this contradicts (3.31). □

Obviously, as $F \equiv -1$ does not solve (3.17) for any finite π_0, it cannot be interpreted as a possible equilibrium configuration of our cap. However, $F \equiv -1$ does correspond to a "flat" pseudoconfiguration, in a sense that we will now make precise.

We begin by noting that, for $F \equiv -1$, we have from (2.30) and $(2.68)_1$ that

$$u^0_\Theta(\Theta) - u^{0\prime}_R(\Theta) = -R_0\Theta. \tag{3.32}$$

We also have from the second equation in Problem $\mathscr{P}'$ together with $(2.68)_2$ and (3.24) that

$$S(\Theta) = \pi_0 - \frac{\Theta_0^2 - \Theta^2}{16}. \tag{3.33}$$

On the other hand, it is not difficult to see that, from a purely kinematic point of view, an axisymmetric flattening displacement of the spherical surface coinciding with the middle surface of the cap in the reference configuration has to have the following form:

$$\begin{aligned} u^0_R(\Theta) &= -R_0(1 - \cos\Theta_0 \cos\Theta) + \sin\Theta v(\Theta), \\ u^0_\Theta(\Theta) &= -R_0 \cos\Theta_0 \sin\Theta + \cos\Theta v(\Theta), \end{aligned} \tag{3.34}$$

where v is an arbitrary (smooth and) odd function. Equivalently, (3.34) can be written as follows:

$$u^0_\Theta(\Theta) - u^{0\prime}_R(\Theta) = -\sin\Theta v'(\Theta), \qquad u^0_R(\Theta) + \cot\Theta u^0_\Theta = -R_0 + \frac{v(\Theta)}{\sin\Theta}, \tag{3.35}$$

or rather, with the use of the shallow cap approximation introduced in § 2.6:

$$u^0_\Theta(\Theta)-u^{0\prime}_R(\Theta)=-\Theta v'(0),\qquad \Theta u^0_R(\Theta)+u^0_\Theta(\Theta)=\Theta(-R_0+v'(0)). \tag{3.36}$$

Comparison of (3.32) and $(3.36)_1$ suggests that we exploit the arbitrariness in the choice of v by taking $v'(0)=R_0$, so that (3.32) is satisfied and $(3.36)_2$ becomes

$$\Theta u^0_R(\Theta)+u^0_\Theta(\Theta)=0. \tag{3.37}$$

But, it follows from (2.32), $(2.51)_{1,2}$, and (2.57) that

$$\frac{1}{R_0}(u^0_R+\cot\Theta u^0_\Theta)=\frac{1}{R_0}(u^0_R+u^{0\prime}_\Theta)+\frac{1}{2R_0^2}(u^0_\Theta-u^{0\prime}_R)^2+(\tan\Theta S)'-S. \tag{3.38}$$

In the shallow cap approximation, (3.37) results from (3.38) by the use of (3.32) and (3.33). We then have reason to say that $F\equiv -1$ corresponds to "flattening" a spherical cap of small, but nonnegligible thickness into a circular disk by means of extremely high outward peripheral tractions.

3.4. Local minimizers. We begin with a lemma that will allow us to define the Leray-Schauder degree of Γ', the Fréchet derivative of the energy functional.

LEMMA 6. *Let* $\{\kappa_{0,n}\}$ *with* $\inf\kappa_{0,n}>0$, *and* $\{\pi_{0,n}\}$ *be two bounded sequences in* $\mathbb{R}$; *moreover, for* $\{f_n\}$ *a sequence in* K^1, *let*

$$\Gamma'(f_n):=f_n+\kappa^2_{0,n}G_1[(1+f_n)G_0[(2+f_n)f_n]+2\pi_{0,n}(1+f_n)]. \tag{3.39}$$

Then, if the sequence $\{\Gamma'(f_n)\}$ *is bounded, the sequence* $\{f_n\}$ *is also bounded in the* K^1*-norm.*

Proof. With a view toward coming up with a contradiction, and with no loss of generality, we suppose that $\|f_n\|_1\to+\infty$. We may also suppose that

$$\frac{1+f_n}{\|1+f_n\|_1}\to h$$

weakly in K^1, and therefore strongly in K. Now, by (3.39) and the definitions of the scalar products for K^1 and K, we have that

$$(\Gamma'(f_n),1+f_n)_1=(f_n,1+f_n)_1+\kappa^2_{0,n}(2\pi_{0,n}+G_0[(2+f_n)f_n],(1+f_n)^2)_0. \tag{3.40}$$

Dividing (3.40) by $\|1+f_n\|_1^4$ and taking the limit for $n\to\infty$, it follows that $(G_0[h^2],h^2)_0=0$, thereby implying that $h=0$, or rather,

$$\frac{\|1+f_n\|_0}{\|1+f_n\|_1}\to 0.$$

From (3.40) we have that

$$(\Gamma'(f_n),1+f_n)_1\geqq(f_n,1+f_n)_1+2\pi_{0,n}\kappa^2_{0,n}(1,(1+f_n)^2)_0-\kappa^2_{0,n}(G_0[1],(1+f_n)^2)_0,$$

and hence

$$\liminf_{n\to\infty}\frac{(\Gamma'(f_n),1+f_n)_1}{\|1+f_n\|_1}\geqq\lim_{n\to\infty}\|1+f_n\|_1=+\infty,$$

a result incompatible with the hypothesis that the sequence $\{\Gamma'(f_n)\}$ is bounded. □

In view of the preceding lemma we have that the Leray-Schauder degree δ of Γ' is well defined and independent of the values of parameters π_0 and κ_0. Due to the convexity of Γ for π_0 large, we also have that $\delta=1$.

We remark that each extremal F of Γ must be such that $F(\Theta_0) \neq -1$. Indeed, consider the flat manifold

$$\mathcal{M} := \{F \in K^1 \,|\, F(\Theta_0) = -1\}$$

and let

$$F \mapsto \tilde{F} := -2 - F$$

be the symmetry mapping of K^1 onto itself about the point $F_0 \equiv -1$. By direct inspection of the definition (3.18) of Γ, we see that $\Gamma|_{\mathcal{M}}$, the restriction of Γ to $\mathcal{M}$, attains its absolute minimum at F_0 and, moreover, for each $F \in \mathcal{M}$, $\Gamma|_{\mathcal{M}}(F) = \Gamma|_{\mathcal{M}}(\tilde{F})$. Suppose absurdly that $F \neq F_0$, and therefore also that $\tilde{F}$ were an extremal of $\Gamma|_{\mathcal{M}}$. Then, the fourth-order polynomial

$$\tau \mapsto \Gamma|_{\mathcal{M}}(-1 + \tau(1+F))$$

would need to have extremals at $\tau = \pm 1$, absolute minimum at $\tau = 0$, and would have to grow to $+\infty$ for $\tau \to \pm\infty$, which is manifestly impossible.

If we denote by $\Omega^+(\Omega^-)$ the collection of all functions $\hat{F}$ in K^1 such that $\hat{F}(\Theta_0) > -1 (\hat{F}(\Theta_0) < -1)$, and with δ^+ and, respectively, δ^- the corresponding degrees, we deduce from the above that both $\delta^+ = 1$ and $\delta^- = 0$.

We now turn to investigating the behavior of local minimizers for the thickness parameter tending to infinity, i.e., when the cap is made thinner and thinner.

For π_0 fixed, let $\{\kappa_{0,n}\}$ be a sequence in $\mathbb{R}$ such that $\{\kappa_{0,n}\} \to +\infty$, and let $\{\hat{F}_n\}$ be a sequence in K^1 whose typical element $\hat{F}_n$ is the extremal of Γ corresponding to (π_0 and) $\kappa_{0,n}$. We will call such a sequence a κ_0-sequence of extremals.

LEMMA 7. *For π_0 fixed, let $\{\hat{F}_n\}$ be a bounded κ_0-sequence of extremals. Then, we have the following*:

(i) *For $\pi_0 \neq 0$, $\{\hat{F}_n\} \to -1$ weakly*;

(ii) *For $\pi_0 = 0$, every weakly convergent subsequence of $\{\hat{F}_n\}$ tends to -2, -1, or zero.*

Proof. Let $\{\hat{F}_n\} \to \hat{F} \in K^1$ weakly. From (3.17), in the limit for $n \to \infty$, we get

$$G_1[(1+\hat{F})G_0[(2+\hat{F})\hat{F}] + 2\pi_0(1+\hat{F})] = 0,$$

and hence

$$(1+\hat{F})(G_0[(2+\hat{F})\hat{F}] + 2\pi_0) = 0. \tag{3.41}$$

On setting

$$\hat{G} := G_0[(2+\hat{F})\hat{F}], \tag{3.42}$$

we obtain from (3.41) and (3.42) that

$$(1+\hat{F})(\hat{G} + 2\pi_0) = 0 \quad \text{and} \quad L[\hat{G}] = (2+\hat{F})\hat{F}. \tag{3.43}$$

Let J denote any connected component of the set $\{\Theta \in]0, \Theta_0[\,|\, \hat{F}(\Theta) \neq -1\}$. As the function $(\hat{G} + 2\pi_0)$ vanishes identically over J, so does the function $(2+\hat{F})\hat{F}$ and, by (3.42), so does also $\hat{G}$. The two assertions under (i) and (ii) easily follow. □

It is interesting to note that the "flat" pseudoconfiguration corresponding to $\hat{F} \equiv -1$ is the common limit of situations when either κ_0 is fixed and the applied load is made to tend to $+\infty$ (Theorem 5) or π_0 is fixed and the thickness is made to tend to zero. For $\pi_0 = 0$, the two other possible weak limits of bounded κ_0-sequences of extremals are the reference configuration, which obtains for $\hat{F} \equiv 0$, and an "everted" pseudoconfiguration corresponding to $\hat{F} \equiv -2$.

To explain the terminology alluding to eversion, we first observe that, for $\hat{F} \equiv -2$, (3.32) and (3.33) are replaced, respectively, by

$$u^0_\Theta - u^{0\prime}_R = -2R_0\Theta \tag{3.44}$$

and

$$S \equiv 0. \tag{3.45}$$

Second, it follows from (2.32), $(2.51)_{1,2}$, (2.57), and (3.1) that

$$u^0_R + \cot\Theta u^0_\Theta = \frac{R_0}{1+\nu}((\tan\Theta S)' - \nu S), \tag{3.46}$$

so that, in view of (3.45), we have that

$$u^0_R + \cot\Theta u^0_\Theta = 0. \tag{3.47}$$

Third, we note the form that a reflection with respect to the horizontal plane of the surface coinciding with the middle surface of the cap in the reference configuration must have

$$u^0_R(\Theta) = -2R_0\cos\Theta(\cos\Theta - \cos\Theta_0), \qquad u^0_\Theta(\Theta) = 2R_0\sin\Theta(\cos\Theta - \cos\Theta_0). \tag{3.48}$$

Equivalently, in such a reflection, (3.47) and

$$u^0_\Theta - u^{0\prime}_R = -2R_0\sin\Theta\cos\Theta$$

have to prevail. In the shallow cap limit we are then entitled to interpret $F \equiv -2$ as the limit situation corresponding to perform an eversion process, at zero loads, of thinner and thinner caps.

We remark that, for $\pi_0 < \Theta_0^2/16$ and κ_0 large, there exists a neighborhood of -1 in K^1 (independent of κ_0) where Γ'' is not semidefinite; therefore, a sequence of local minimizers cannot converge to -1. Thus, in view of Lemma 7, if $\pi_0 = 0$, a κ_0-sequence of local minimizers breaks down into subsequences converging weakly to $\hat{F} \equiv 0$ or $\hat{F} \equiv -2$, or else diverging in the K^1-norm; if $\pi_0 \neq 0$, a κ_0-sequence of local minimizers has to diverge.

To put the last remark into perspective, we observe that

$$\Gamma(F_n) \leqq \Gamma(0) = 0$$

for $\{F_n\}$ any sequence of global minimizers. With a glance to (3.18) we see that, for $\pi_0 > 0$, we must have

$$(2 + F_n, F_n)_0 \leqq 0,$$

or rather

$$\|1 + F_n\|_0^2 < \|1\|_0^2.$$

Thus, when $\pi_0 > 0$, a sequence of global minimizers cannot diverge in the K-norm.

We now show that local minimizers do exist in Ω^-.

THEOREM 8. *Let $\pi_0 = 0$. Then, there exists $\bar{\kappa}_0 > 0$ such that, for all $\kappa_0 > \bar{\kappa}_0$, the energy functional Γ has a local minimizer $\hat{F}_{\kappa_0}$ belonging to Ω^-. Moreover, a κ_0-sequence of such local minimizers tends to $F \equiv -2$.*

Proof. Choose $\gamma_1 > \|-2\|_1$ and $\gamma_2 \in]-2, -1[$, and define

$$\Delta := \{f \in K^1 \mid \|f\|_1 \leqq \gamma_1, f(\Theta_0) \leqq \gamma_2\},$$

with $\partial\Delta$, the boundary of Δ, such that

$$\partial\Delta = \partial_1\Delta \cup \partial_2\Delta \quad \text{with} \quad \begin{aligned} \partial_1\Delta &:= \{f \in \partial\Delta \,|\, \|f\| = \gamma_1\}, \\ \partial_2\Delta &:= \{f \in \partial\Delta \,|\, f(\Theta_0) = \gamma_2\}. \end{aligned}$$

As Δ is a closed and convex set, Γ must attain an absolute minimum on Δ; on the other hand, $F \equiv -2$ is an interior point of Δ itself. We will now establish the first statement of the theorem by proving that there exists $\bar{\kappa}_0(\gamma_1, \gamma_2)$ such that

$$\inf_{F \in \partial\Delta} \Gamma(F) > \Gamma(-2) \quad \text{for } \kappa_0 > \bar{\kappa}_0(\gamma_1, \gamma_2). \tag{3.49}$$

Indeed, for $F \in \partial_1\Delta$,

$$\Gamma(F) \geqq \tfrac{1}{2}\gamma_1^2 > \Gamma(-2).$$

Furthermore, for $F \in \partial_2\Delta$, let

$$\gamma_3 := \min_{F \in \partial_2\Delta} \|G_0[(2+F)F]\|_1^2 > 0$$

(as $\partial_2\Delta$ is a closed and convex set, this definition makes sense). Then, it is always possible to choose $\bar{\kappa}_0(\gamma_1, \gamma_2)$ such that, for $\bar{\kappa}_0 < \kappa_0$,

$$\Gamma(F) \geqq \frac{\kappa_0^2}{4}\gamma_3 > \Gamma(-2).$$

Finally, the second statement of the theorem follows from the fact that the Δ sets form a neighborhood base for $F \equiv -2$. □

In the light of this theorem, among the three weak limits possible for bounded κ_0-sequences of extremals according to Lemma 7 when $\pi_0 = 0$, the "everted" pseudoconfiguration $\hat{F} \equiv -2$ is the only one that is a local minimizer.

In Fig. 2, for $\pi_0 = 0$, $\nu = 0.3$, and $\Theta_0 = 0.4$ rad, the everted configurations have been graphed that are obtained by solving system (2.67) with the boundary conditions (2.69) for various values of the thickness parameter κ_0.

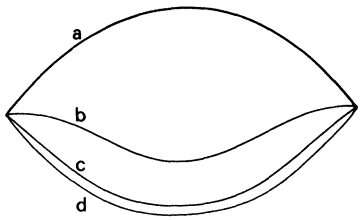


FIG. 2. *Reference configuration* (a) *and everted shapes for* $\pi_0 = 0$, $\Theta_0 = 0.4$, $\nu = 0.3$, *and* $\kappa_0 = 133$ (b), 313 (c), 1875 (d).

For relatively small κ_0 (a relatively thick cap), an edge flaring effect is evident. We can also see that the everted shape increasingly resembles the reflected shape of the reference configuration as κ_0 increases. For completeness, the (rescaled) bending moment M and stress S corresponding to case b of Fig. 2 are graphed in Fig. 3.

We remark that, as the degree of Γ' in Ω^- is null and as a local minimizer exists, another extremal of Γ must be found in Ω^-. We have been unable to spot this other extremal, either analytically or numerically.

Finally, we remark that, as inequality (3.49) holds true also for small values of π_0, Theorem 8 can be established also under the weaker hypothesis that $|\pi_0| \ll 1$.

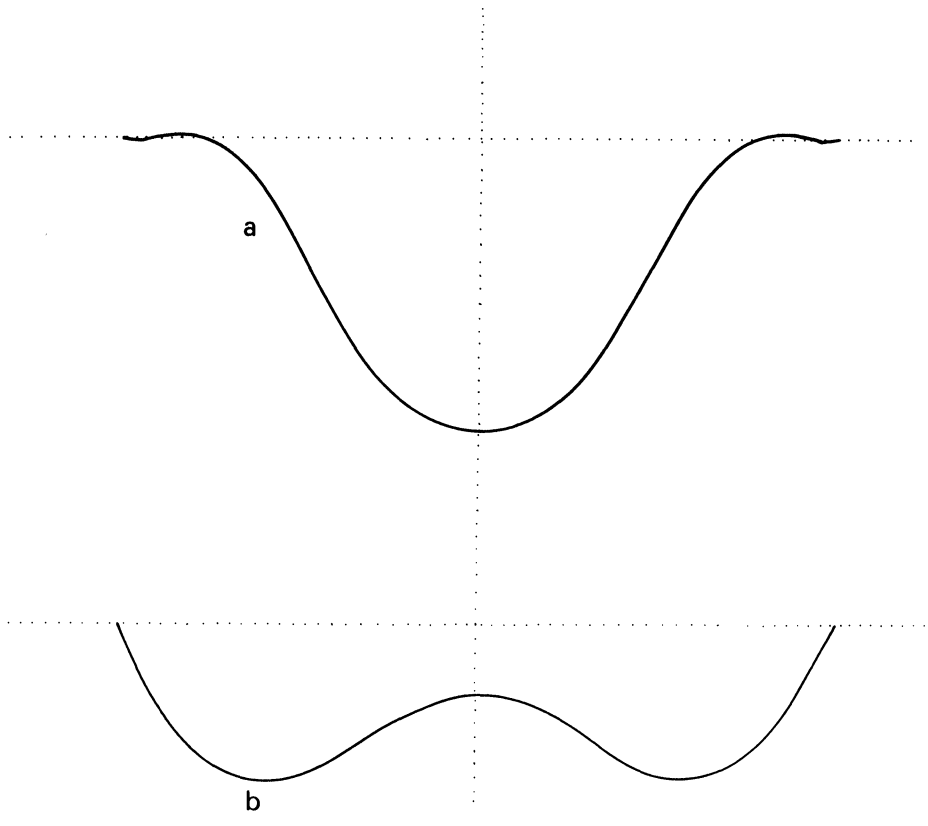


FIG. 3. *Moment* (a, $10^{-1} \times M$) *and stress* (b, $50 \times S$) *for the everted configuration* b *of Fig.* 2.

REFERENCES

[1] C. TRUESDELL, *Some challenges offered to analysis by rational thermodynamics*, in Contemporary Developments in Continuum Mechanics and Partial Differential Equations, G. M. de La Pehha and L. A. Medeiros, eds., North-Holland, Amsterdam, 1978.

[2] S. A. ANTMAN, *The eversion of thick spherical shells*, Arch. Rational Mech. Anal., 70 (1979), pp. 113-123.

[3] P. G. CIARLET AND G. GEYMONAT, *Sur le lois de comportement en élasticité non linéaire compressible*, C.R. Acad. Sci. Paris, 295 (1982), pp. 423-426. (A thorough discussion of Saint-Venant & Kirchhoff materials is contained in Mathematical Elasticity, Vol. I: Three-Dimensional Elasticity, P. G. Ciarlet, North-Holland, Amsterdam, 1988.)

[4] G. GEYMONAT, M. ROSATI, AND V. VALENTE, *Numerical analysis for eversion of elastic spherical caps*, in Proc. 8th Internat. Conference on Computing Methods in Applied Sciences and Engineering, Versailles, December 14-18, 1987.

[5] A. E. H. LOVE, *A Treatise on the Mathematical Theory of Elasticity*, Cambridge University Press, Cambridge, 1927.

[6] S. TIMOSHENKO AND S. WOINOWSKI-KRIEGER, *Theory of Plates and Shells*, McGraw-Hill, New York, 1982.

[7] M. DIKMEN, *Theory of Thin Elastic Shells*, Pitman, Boston, 1982.

[8] V. V. NOVOZHILOV, *Foundations of the Nonlinear Theory of Elasticity*, Graylock Press, 1953.

[9] P. G. CIARLET AND J. C. PAUMIER, *A justification of the Marguerre-von Karman equations*, Comput. Mech., 1 (1986), pp. 1771-202.

[10] K. Z. GALIMOV AND KH. M. MUSTARI, *Non Linear Theory of Thin Elastic Shells*, S. Monson, Jerusalem, 1961.

[11] E. L. REISS, *Bifurcation buckling of spherical caps*, Comm. Pure Appl. Math., 17 (1965), pp. 65-82.
[12] L. BAUER, E. L. REISS, AND H. B. KELLER, *Axisymmetric buckling of hollow spheres and hemispheres*, Comm. Pure Appl. Math., 23 (1970), pp. 529-568.
[13] S. N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, New York, 1982.
[14] M. A. KRASNOSEL'SKII, *Topological Methods in the Theory of Nonlinear Integral Equations*, Internat. Series Monographs Pure and Applied Mathematics, 45, Pergamon Press, Elmsford, N.Y., 1964.
[15] J. T. SCHWARTZ, *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 664–700, May 1989

012

$M(\lambda)$ THEORY FOR SINGULAR HAMILTONIAN SYSTEMS WITH ONE SINGULAR POINT*

ALLAN M. KRALL†

Abstract. The theory of singular Hamiltonian systems is developed. Square integrable solutions are exhibited and used to define Green's function. Using a singular Green's formula, other self-adjoint boundary value problems are generated in which regular and singular boundary conditions are mixed together. Finally the spectral measure, the generalized Fourier transform of an arbitrary function, and the inverse transform for problems with separated boundary conditions are derived.

Key words. singular boundary value problems, Sturm–Liouville problem, spectral resolution

AMS(MOS) subject classifications. 34B05, 34B20, 34B25

1. Introduction. The solution of linear boundary value problems has a long and honored history. Rising from attacks on such problems as the solution of the heat, wave, and Laplace equation as well as others from mathematical physics, linear boundary value problems have played an important part in mathematics for over two centuries.

The problems fall into two classifications. First, those defined over finite intervals with well-behaved coefficients are called regular. They have a discrete set of eigenvalue, eigenfunction pairs, which are used as building blocks in the solution of the partial differential equations from mathematical physics. The solution of such problems is classical and may be found in many books, such as [3].

Problems that are not regular are singular. These are considerably more difficult to discuss, and as a result have only been examined closely during this century. The work was begun by Weyl [32] in 1910. He was followed by Titchmarsh [30] and many others. From 1910 until 1945 these mathematicians developed and polished the theory of self-adjoint differential operators of second order to a high degree.

Their work was continued by Kodaira [21], Coddington and Levinson [3], and Hartman and Wintner (see [4]) in the late 1940s and 1950s. Not only were additional results found for operators of second order, but operators of higher order were also examined. At the same time the Russian school, led by Krein, Naimark, Akhiezer, and Glazman, also made major contributions.

For a far more comprehensive survey of this work, we recommend the second volume of Dunford and Schwartz [4]. They provide an excellent summary of the numerous contributions made by many mathematicians.

Further study continued in the 1960s and 1970s with the work of Atkinson [1] on regular Hamiltonian systems and Everitt [5]–[11] on higher-order scalar problems. The work of this period is summarized by Atkinson [1], Everitt and Kumar [12] and by Kogan and Rofe-Beketov [22] of the Russian school. Again, there were many other contributors. One contribution, perhaps, deserves special mention. Walker [31] showed that any scalar self-adjoint problem of an arbitrary order can be reformulated as an equivalent self-adjoint Hamiltonian system. This removed the need to discuss scalar problems and systems separately.

* Received by the editors May 17, 1987; accepted for publication (in revised form) August 9, 1988. This work was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, contract W-31-109-Eng-38.

† Department of Mathematics, McAllister Building, Pennsylvania State University, University Park, Pennsylvania 16802.

Most recently in the 1980s Hinton and Shaw [13]-[20] have made great progress by considering singular problems in the Hamiltonian system format, following the lead of Atkinson [1]. Many of the results of the past, including the development of $M(\lambda)$ theory, the derivation of the Green's function, and singular boundary conditions, were substantially improved.

We will follow in the footsteps of Atkinson, Hinton, and Shaw by considering the $2n$-dimensional linear Hamiltonian system $JY' = (\lambda A + B)Y$, over an interval $[a, b)$. The procedure we will follow is to first examine the differential equation over a subinterval $[a, b']$, $a < b' < b$, on which the problem is regular. We then allow b' to approach b.

The first problem to be encountered concerns the number of solutions that lie in the underlying Hilbert space $L^2_A(a, b)$, generated by the inner product

$$\langle Y, Z\rangle = \int_a^b Z^* A Y \, dt.$$

For $\operatorname{Im} \lambda \neq 0$ it is possible to show that there exist m solutions in $L^2_A(a, b)$, $n \leqq m \leqq 2n$. Since, in order to exhibit a Green's function, only n are needed, the problem is to determine which to choose.

There are two easier to handle situations. If the number is n, then there is no freedom. Everything is determined. If the number is $2n$, then there are automatically a number of required limits. But a number of problems exist in the situations involving m solutions in $L^2_A(a, b)$, $n < m < 2n$. These are handled by the use of singular boundary conditions. One main purpose of this paper is to show how this is done.

Another problem arises if we attempt to impose boundary conditions arbitrarily. When is a problem self-adjoint? We will show how to use Green's formula to determine all self-adjoint problems. We will do so in a manner that considerably simplifies previously used approaches [4].

Finally, as we allow b' to approach b, the generalized eigenfunction expansion over $[a, b']$ is shown to converge into the general spectral resolution traditionally associated with self-adjoint operators. The generalized Fourier transform and inverse transform will be explicitly exhibited.

2. Notation and definition. We consider over the interval $[a, b)$ the differential expression

$$(*) \qquad JY' = (\lambda A + B)Y,$$

where Y is of dimension $2n$ (a $2n \times 1$ matrix).

$$J = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}, \qquad A = \begin{pmatrix} A_{11}(x) & A_{12}(x) \\ A_{21}(x) & A_{22}(x) \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11}(x) & B_{12}(x) \\ B_{21}(x) & B_{22}(x) \end{pmatrix}$$

are locally integrable $2n \times 2n$ matrices, $A = A^* \geqq 0$ and $B = B^*$. We assume that a is a regular point, while b is singular. That is, a is finite; A and B are integrable in a neighborhood of a. b may be finite or infinite; A and B may not be integrable in a neighborhood of b. The word singular is used to denote difficulties with infinity or with nonintegrability.

Our setting is $L^2_A(a, b)$, the Hilbert space generated by the inner product

$$\langle Y, Z\rangle = \int_a^b Z^* A Y \, dt.$$

To ensure that elements in the domain of the maximal operator, to be defined, are dense in $L_A^2(a, b)$, we assume that if $JY' - BY = AF$ and $AY = 0$, then $Y = 0$. Most of what follows proceeds without this assumption, but without it some expansions must be restricted to subspaces instead of holding on all of $L_A^2(a, b)$. Earlier works ([1], [13]-[29]) made this assumption with $F = 0$.

We impose a regular, self-adjoint boundary condition at a,

$$(\alpha_1, \alpha_2)\, Y(a) = 0,$$

where α_1, α_2 are $n \times n$ matrices satisfying rank $(\alpha_1, \alpha_2) = n$ and

$$\alpha_1 \alpha_1^* + \alpha_2 \alpha_2^* = I, \qquad \alpha_1 \alpha_2^* - \alpha_2 \alpha_1^* = 0.$$

Self-adjointness is assured, since when written as $MY(a) = 0$, where

$$M = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 0 & 0 \end{pmatrix},$$

$$MJM^* = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} \alpha_1^* & 0 \\ \alpha_2^* & 0 \end{pmatrix} = 0,$$

the requirement for self-adjointness under separated conditions [3, p. 291].

It is no imposition to assure that $\alpha_1 \alpha_1^* + \alpha_2 \alpha_2^* = I_n$, for if the rank $(\alpha_1, \alpha_2) = n$, then

$$\operatorname{rank} \begin{pmatrix} \alpha_1^* \\ \alpha_2^* \end{pmatrix} = n,$$

and the rank of

$$\alpha_1 \alpha_1^* + \alpha_2 \alpha_2^* = (\alpha_1 \alpha_2) \begin{pmatrix} \alpha_1^* \\ \alpha_2^* \end{pmatrix}$$

is n. It is nonsingular and positive. Hence if the sum does not equal I_n, replace α_1 and α_2 by $(\alpha_1 \alpha_1^* + \alpha_2 \alpha_2^*)^{1/2} \alpha_1$ and $(\alpha_1 \alpha_1^* + \alpha_2 \alpha_2^*)^{1/2} \alpha_2$.

Next let b' be in (a, b) and impose the regular, self-adjoint boundary condition

$$(\beta_1, \beta_2)\, Y(b') = 0,$$

where

$$\beta_1 \beta_1^* + \beta_2 \beta_2^* = I_n, \qquad \beta_1 \beta_2^* + \beta_2 \beta_1^* = 0.$$

If the boundary condition is written as $NY(b') = 0$, where

$$N = \begin{pmatrix} 0 & 0 \\ \beta_1 & \beta_2 \end{pmatrix},$$

then

$$NJN^* = \begin{pmatrix} 0 & 0 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} 0 & \beta_1^* \\ 0 & \beta_2^* \end{pmatrix} = 0.$$

Again the requirement for self-adjointness is under separated conditions.

Problem (*), together with the two boundary conditions, defines a regular, self-adjoint Sturm-Liouville problem.

Finally, let

$$E = \begin{pmatrix} \alpha_1^* & -\alpha_2^* \\ \alpha_2^* & \alpha_1^* \end{pmatrix}$$

and let $\mathcal{Y}$ be the fundamental matrix for (*) satisfying $\mathcal{Y}(a) = E$.

If $\mathcal{Y}$ is partitioned into

$$\mathcal{Y} = (\theta, \phi) = \begin{pmatrix} \theta_1 & \phi_1 \\ \theta_2 & \phi_2 \end{pmatrix},$$

then at $x = a$, $(\alpha_1, \alpha_2)\theta(a) = I_n$, and $(\alpha_1, \alpha_2)\phi(a) = 0$. $\phi = \binom{\phi_1}{\phi_2}$ satisfies the boundary condition at a.

3. The $M(\lambda)$ matrix. The following definition and characterization of the matrix $M(\lambda)$ is a generalization of the Weyl-Titchmarsh $m(\lambda)$ function [30], [31]. What immediately follows is closely patterned on the work of Hinton and Shaw [13]-[20]. It is essential in providing the building blocks needed throughout the remainder of the paper: the square integrable solutions of (*).

THEOREM 3.1. *Let β_1, β_2 satisfy*

$$\beta_1\beta_1^* + \beta_2\beta_2^* = I_n, \qquad \beta_1\beta_2^* - \beta_2\beta_1^* = 0.$$

Let

$$\chi_{b'} = \mathcal{Y}\begin{pmatrix} I_n \\ M(b') \end{pmatrix},$$

and suppose $(\beta_1, \beta_2)\chi_{b'}(b') = 0$. Then

$$M(b') = -(\beta_1\phi_1(b') + \beta_2\phi_2(b'))^{-1}(\beta_1\theta_1(b') + \beta_2\theta_2(b'))$$

and $\chi_{b'}^*(b')J\chi_{b'}(b') = 0$.

Conversely, if for some M,

$$\chi_{b'} = \mathcal{Y}\begin{pmatrix} I_n \\ M \end{pmatrix}$$

satisfies $\chi_{b'}^*(b')J\chi_{b'}(b') = 0$, then there exist β_1, β_2 satisfying

$$\beta_1\beta_1^* + \beta_2\beta_2^* = I_n, \qquad \beta_1\beta_2^* - \beta_2\beta_1^* = 0$$

such that $(\beta_1, \beta_2)\chi_{b'}(b') = 0$ and

$$M = -(\beta_1\phi_1(b') + \beta_2\phi_2(b'))^{-1}(\beta_1\theta_1(b') + \beta_2\theta_2(b')).$$

Proof. Let $\operatorname{Im}\lambda \neq 0$, and impose on

$$\chi_{b'} = \mathcal{Y}\begin{pmatrix} I_n \\ M(b') \end{pmatrix}$$

the boundary condition $(\beta_1, \beta_2)\chi_{b'}(b') = 0$. Hence, with $x = b'$,

$$(\beta_1, \beta_2)\begin{pmatrix} \theta_1 & \phi_1 \\ \theta_2 & \phi_2 \end{pmatrix}\begin{pmatrix} I_n \\ M(b') \end{pmatrix} = 0.$$

This yields

$$(\beta_1\theta_1(b') + \beta_2\theta_2(b')) + (\beta_1\phi_1(b') + \beta_2\phi_2(b'))M(b') = 0$$

and

$$M(b') = -(\beta_1\phi_1(b') + \beta_2\phi_2(b'))^{-1}(\beta_1\theta_1(b') + \beta_2\theta_2(b')).$$

The inverse must exist. Otherwise λ, which is complex, would be an eigenvalue of the self-adjoint boundary value problem on $[a, b']$.

Since $(\beta_1, \beta_2)\chi_{b'}(b') = 0$,

$$\chi(b') = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} C,$$

for

$$(\beta_1, \beta_2) \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} C = 0.$$

This in turn implies that $\chi_{b'}^*(b') J \chi_{b'}(b') = 0$, or

$$(I_n, M(b')^*) \mathcal{Y}(b')^* J \mathcal{Y}(b') \begin{pmatrix} I_n \\ M(b') \end{pmatrix} = 0.$$

Conversely, if for some M,

$$(I_n, M^*) \mathcal{Y}(b')^* J \mathcal{Y}(b') \begin{pmatrix} I_n \\ M \end{pmatrix} = 0,$$

define $(\beta_1, \beta_2) = (I_n, M^*) \mathcal{Y}(b')^* J$. Then rank $(\beta_1, \beta_2) = n$,

$$(\beta_1, \beta_2) \mathcal{Y}(b') \begin{pmatrix} I_n \\ M \end{pmatrix} = 0,$$

and

$$-\beta_1 \beta_2^* + \beta_2 \beta_1^* = (\beta_1, \beta_2) \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = (I, M^*) \mathcal{Y}(b')^* J^3 \mathcal{Y}(b') \begin{pmatrix} I \\ M \end{pmatrix} = 0.$$

Further $M = -(\beta_1 \phi_1(b') + \beta_2 \phi_2(b'))^{-1} (\beta_1 \theta_1(b') + \beta_2 \phi_2(b'))$, and

$$\beta_1 \beta_1^* + \beta_2 \beta_2^* = (\beta_1, \beta_2) \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} = (I_n, M^*) \mathcal{Y}(b')^* \mathcal{Y}(b') \begin{pmatrix} I_n \\ M \end{pmatrix} > 0,$$

and so with minor adjustments like those of the previous section, we can have $\beta_1 \beta_1^* + \beta_2 \beta_2^* = I_n$.

4. M circles. The M circle equation is

$$\pm (I_n, M^*) \mathcal{Y}(b')^* (J/i) \mathcal{Y}(b') \begin{pmatrix} I_n \\ M \end{pmatrix} = 0$$

where for convenience, we have divided by i, and $(+)$ holds when $\operatorname{Im} \lambda > 0$, $(-)$ holds when $\operatorname{Im} \lambda < 0$.

Let

$$\begin{pmatrix} \mathcal{A} & \mathcal{B}^* \\ \mathcal{B} & \mathcal{D} \end{pmatrix} = \begin{cases} \mathcal{Y}(b')^* (J/i) \mathcal{Y}(b'), & \operatorname{Im} \lambda > 0, \\ -\mathcal{Y}(b')^* (J/i) \mathcal{Y}(b'), & \operatorname{Im} \lambda < 0, \end{cases}$$

and let

$$E(M) = (I_n, M^*) \begin{pmatrix} \mathcal{A} & \mathcal{B}^* \\ \mathcal{B} & \mathcal{D} \end{pmatrix} \begin{pmatrix} I_n \\ M \end{pmatrix}.$$

THEOREM 4.1. *The expression $E(M) = 0$ if and only if $M = M_\beta$ for some $\beta = (\beta_1, \beta_2)$, where*

$$M_\beta = -(\beta_1 \phi_1(b') + \beta_2 \phi_2(b'))^{-1} (\beta_1 \theta_1(b') + \beta_2 \theta_2(b')).$$

This is a restatement of Theorem 3.1.

Expanding, we find

$$\begin{aligned} E(M) &= (M+\mathscr{D}^{-1}\mathscr{B})^*\mathscr{D}(M+\mathscr{D}^{-1}\mathscr{B})+\mathscr{A}-\mathscr{B}^*\mathscr{D}^{-1}\mathscr{B} \\ &= (M-C)^*R_1^{-2}(M-C)-R_2^2 \\ &= 0, \end{aligned}$$

where $C=-\mathscr{D}^{-1}\mathscr{B}$, $R_1=\mathscr{D}^{-1/2}$, $R_2=(\mathscr{B}^*\mathscr{D}^{-1}\mathscr{B}-\mathscr{A})^{1/2}$.

LEMMA 4.2. *The matrix* $\mathscr{D}>0$.

Proof.

$$\begin{pmatrix} \mathscr{A} & \mathscr{B}^* \\ \mathscr{B} & \mathscr{D} \end{pmatrix} = \pm \begin{pmatrix} \theta_1^* & \theta_2^* \\ \phi_1^* & \phi_2^* \end{pmatrix} \begin{pmatrix} 0 & iI_n \\ -iI_n & 0 \end{pmatrix} \begin{pmatrix} \theta_1 & \phi_1 \\ \theta_2 & \phi_2 \end{pmatrix} = \pm \begin{pmatrix} -i\theta^*J\theta & -i\theta^*J\phi \\ -i\phi^*J\theta & -i\phi^*J\phi \end{pmatrix}$$

where

$$\phi = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \qquad \phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}.$$

So $\mathscr{D}=\pm\phi^*(b')(J/i)\phi(b')$. Now manipulation of the differential equation $J\phi'=(\lambda A+B)\phi$ yields

$$\phi^*(J/i)\phi|_a^{b'} = 2 \operatorname{Im}\lambda \int_a^{b'} \phi^* A\phi \, dt.$$

The limit at $x=a$ is 0. Hence $\phi(b')^*(J/i)\phi(b')>0$ when $\operatorname{Im}\lambda>0$, and $\phi(b')^*(J/i)\phi(b')<0$ when $\operatorname{Im}\lambda<0$. In either case $\mathscr{D}>0$.

LEMMA 4.3. *The matrix* $\mathscr{B}^*\mathscr{D}\mathscr{B}-\mathscr{A}=\tilde{\mathscr{D}}^{-1}$, *where* $\tilde{\mathscr{D}}^{-1}=\mathscr{D}^{-1}(\bar{\lambda})$.

Proof. Note that for λ and $\bar{\lambda}$, $\mathscr{Y}(x,\bar{\lambda})^*J\mathscr{Y}(x,\lambda)=J$ for all x. Hence $-J\mathscr{Y}(x,\bar{\lambda})^*J\mathscr{Y}(x,\lambda)=I$ and $(J\mathscr{Y}(x,\lambda))(-J\mathscr{Y}(x,\bar{\lambda})^*)=I$ as well. Multiplying by J yields $\mathscr{Y}(x,\lambda)J\mathscr{Y}(x,\bar{\lambda})^*=J$ for all x. As a result

$$\begin{aligned} J &= \mathscr{Y}(x,\lambda)^*J\mathscr{Y}(x,\bar{\lambda}) \\ &= \mathscr{Y}(x,\lambda)^*[-J\mathscr{Y}(x,\lambda)J\mathscr{Y}(x,\bar{\lambda})^*J]\mathscr{Y}(x,\bar{\lambda}) \\ &= -[\mathscr{Y}(x,\lambda)^*(J/i)\mathscr{Y}(x,\lambda)]J[-\mathscr{Y}(x,\bar{\lambda})^*(J/i)\mathscr{Y}(x,\bar{\lambda})]. \end{aligned}$$

Or

$$\begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} = -\begin{pmatrix} \mathscr{A} & \mathscr{B}^* \\ \mathscr{B} & \mathscr{D} \end{pmatrix} \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} \tilde{\mathscr{A}} & \tilde{\mathscr{B}}^* \\ \tilde{\mathscr{B}} & \tilde{\mathscr{D}} \end{pmatrix}$$

where $\tilde{}$ indicates $\bar{\lambda}$ replaces λ. (Remember that there is a sign change in the matrix when λ replaces $\bar{\lambda}$.) Hence

$$0=\mathscr{A}\tilde{\mathscr{B}}-\mathscr{B}^*\tilde{\mathscr{A}}, \qquad -I_n=\mathscr{A}\tilde{\mathscr{D}}-\mathscr{B}^*\tilde{\mathscr{B}},$$

$$I_n=\mathscr{B}\tilde{\mathscr{B}}-\mathscr{D}\tilde{\mathscr{A}}, \qquad 0=\mathscr{B}\tilde{\mathscr{D}}-\mathscr{D}\tilde{\mathscr{B}}^*.$$

The last yields $\tilde{\mathscr{B}}^*\tilde{\mathscr{D}}^{-1}=\mathscr{D}^{-1}\mathscr{B}$. The second shows

$$\begin{aligned} \tilde{\mathscr{D}}^{-1} &= -\mathscr{A}+\mathscr{B}^*[\tilde{\mathscr{B}}^*\tilde{\mathscr{D}}^{-1}] \\ &= -\mathscr{A}+\mathscr{B}^*\mathscr{D}^{-1}\mathscr{B}. \end{aligned}$$

COROLLARY 4.4. *The matrices* R_1 *and* R_2 *satisfy* $R_2=\tilde{R}_1$.

Note that if the coefficient matrices A and B are real, then $\mathcal{Y}(x, \bar{\lambda}) = \overline{\mathcal{Y}(x, \lambda)}$ and so also for $\mathcal{A}, \mathcal{B}, \mathcal{D}$. In this case, $R_2 = \bar{R}_1$.

THEOREM 4.5. *As b' increases, $\mathcal{D}$ increases, R_1 decreases, and R_2 decreases.*

Proof. Note that

$$\mathcal{D} = 2|\operatorname{Im} \lambda| \int_a^{b'} \phi^* A \phi \, dt.$$

The results are then immediate.

THEOREM 4.6. $\lim_{b' \to b} R_1(b', \lambda) = R_0(\lambda)$, $\lim_{b' \to b} R_2(b', \lambda) = R_0(\bar{\lambda}) = \tilde{R}_0$ *exist.* $R_0 \geqq 0$, $\tilde{R}_0 \geqq 0$.

THEOREM 4.7. *As b' approaches b, the circles $E(M) = 0$ are nested.* $\lim_{b' \to b} C(b', \lambda) = C_0$ *exists.*

Proof. The interior of the circle $E(M) = 0$ is given by $E(M) \leqq 0$, or by

$$\pm(I_n, M^*) \mathcal{Y}(b')^* (J/i) \mathcal{Y}(b') \begin{pmatrix} I_n \\ M \end{pmatrix} \leqq 0.$$

Using the differential equation (*), we find

$$E(M) = 2|\operatorname{Im} \lambda| \int_a^{b'} \chi_{b'}^* A \chi_{b'} \, dt \pm (M^* - M)/i.$$

Now if M is in the circle at $b'' > b'$, then $E(M) \leqq 0$ at b''. At b', $E(M)$ is certainly smaller, and so M is in the circle at b' as well. The circles are nested.

To show that the centers converge, we need to solve the circle equation

$$(M - C)^* R_1^{-2} (M - C) = \tilde{R}_1^2.$$

(Recall that $\mathcal{D}$, R_1, and $\tilde{R}_1$ are self-adjoint matrices.) This is equivalent to

$$[R_1^{-1}(M - C)\tilde{R}_1^{-1}]^* [R_1^{-1}(M - C)\tilde{R}_1^{-1}] = I_n.$$

Therefore the bracketed term

$$R_1^{-1}(M - C)\tilde{R}_1^{-1} = U,$$

a unitary matrix, and

$$M = C + R_1 U \tilde{R}_1.$$

As U varies over the $n \times n$ unit sphere, M varies over a "circle" with center C. We will have more to say about the range of M shortly.

Now let C_1 be the center at b', and C_2 be the center at b''. If

$$M_1 = C_1 + R_1(b') U_1 \tilde{R}_1(b')$$

and

$$M_2 = C_2 + R_1(b'') U_2 \tilde{R}_2(b''),$$

then M_2 lies in the b' circle as well, and

$$M_2 = C_1 + R_1(b') V_1 \tilde{R}_1(b')$$

where V_1 is a contraction. Thus

$$C_1 - C_2 = R_1(b'') U_2 \tilde{R}_1(b'') - R_1(b') V_1 \tilde{R}_1(b').$$

Consider the mapping defined by the equations for M_2 above, defining V_1 in terms of U_2: $V_1 = F(U_2)$. It is a continuous transformation of the unit sphere into itself and therefore has a unique fixed point U. Letting U_2 and V_1 be replaced by U, we find

$$\begin{aligned}\|C_1 - C_2\| &= \|R_1(b'')U\tilde{R}_1(b'') - R_1(b')U\tilde{R}_1(b')\| \\ &\leqq \|R_1(b'')U\tilde{R}_1(b'') - R_1(b'')U\tilde{R}_1(b')\| \\ &\quad + \|R_1(b'')U\tilde{R}_1(b') - R_1(b')U\tilde{R}_1(b')\| \\ &\leqq \|R_1(b'')\|\,\|\tilde{R}_1(b'') - \tilde{R}_1(b')\| + \|R_1(b'') - R_1(b')\|\,\|\tilde{R}_1(b')\|.\end{aligned}$$

As b' and b'' approach b, R_1 and $\tilde{R}_1$ have limits. The centers then form a Cauchy sequence and converge.

A computation shows

$$\mathscr{B} = \pm\left[2\,\mathrm{Im}\,\lambda \int_a^{b'} \phi^* A\theta\, dt - iI_n\right].$$

So at b', the center

$$C = -\mathscr{D}^{-1}\mathscr{B} = -\left[2\,\mathrm{Im}\,\lambda \int_a^{b'} \phi^* A\phi\, dt\right]^{-1}\left[2\,\mathrm{Im}\,\lambda \int_a^{b'} \phi^* A\theta\, dt - iI_n\right].$$

As we have seen, as b' approaches b, this has a limit C_0.

The limiting "circle" equation may not exist because both R_0 and $\tilde{R}_0$ may have rank less than n and may be singular. Nonetheless

$$M = C_0 + R_0 U\tilde{R}_0$$

is perfectly well defined. As U varies over the unit circle in $n \times n$ space, the limit "circle" or "point" $E_0(M)$ is covered.

5. Square integrable solutions. As a consequence of the previous section we have the following theorem.

THEOREM 5.1. *Let M be a point inside $E_0(M) \leqq 0$. Let $\chi = \theta + \phi M$; then χ is in $L_A^2(a, b)$.*

Proof.

$$2|\mathrm{Im}\,\lambda| \int_a^{b'} \chi^* A\chi\, dt \pm [M^* - M]/i = E(M) \leqq 0.$$

As a result,

$$0 \leqq \int_a^{b'} \chi^* A\chi\, dt \leqq [M - M^*]/2i|\mathrm{Im}\,\lambda|.$$

The upper bound is fixed, so let $b' \to b$. Remember that

$$M = C_0 + R_0 U\tilde{R}_0$$

where R_0 and $\tilde{R}_0$ are decreasing matrices and U is either unitary or a contraction.

THEOREM 5.2. *Let* rank $R_0 = r$, rank $\tilde{R}_0 = \tilde{r}$; *let $S(U) = R_0 U\tilde{R}_0$ where U is unitary. Then* rank $S(U) \leqq \min(r, \tilde{r})$.

Proof. That the rank of a product of matrices is less than or equal to the ranks of the components may be found in virtually any linear algebra text.

THEOREM 5.3. *Under the conditions of Theorem* 5.2, $\sup_U$ rank $S(U) = \min(r, \tilde{r})$.

Proof. Let $r \geqq \tilde{r}$. Otherwise consider the transpose $\tilde{R}_0^T U^T R_0^T$. The rank of $S(U)$ is the dimension of the image of $R_0 U\tilde{R}_0$ acting on C^n. Let $\tilde{R}_0$ acting on C^n be the

subspace W of dimension $\tilde{r} = \text{rank } \tilde{R}_0$. Further note that there is a subspace X of C^n such that $\dim R_0 X = \text{rank } R_0 = r$. Since $\dim X \geqq \dim W$, there is a unitary matrix $U: W \to X$, injectively. Then $U(W)$ is a subspace of X of dimension $\tilde{r}$, and $\dim (R_0 U(W)) = \dim W = \tilde{r}$.

THEOREM 5.4. *Let $m = n + \min(r, \tilde{r})$; let $\text{Im } \lambda \neq 0$. Then there exist m solutions of (*) in $L_A^2(a, b)$, $n \leqq m \leqq 2n$.*

Proof. $\vec{\theta} + \vec{\phi} C_0$ is made up of n solutions in $L_A^2(a, b)$. As U varies, $\phi(R_0 U \tilde{R}_0)$ gives an additional $m - n$ solutions, which are independent of the others. This does not say the deficiency indices for (*) are equal. It says merely that there are at least m solutions in $L_A^2(a, b)$. For λ or $\bar{\lambda}$ there may be more.

By way of example consider Table 1. These cases actually occur in practice for scalar problems. It is also possible for rank $R_0 \neq$ rank $\tilde{R}_0$. McLeod [33] has found a very interesting example.

TABLE 1

n	rank R_0	rank $\tilde{R}_0$	m	Case name
1	1	1	2	limit circle
1	0	0	1	limit point
2	2	2	4	limit circle
2	1	1	3	intermediate
2	0	0	2	limit point
3	3	3	6	limit circle
3	2	2	5	intermediate
3	1	1	4	intermediate
3	0	0	3	limit point

In closing this section we include a rather interesting theorem concerning the eigenvalues of $\mathcal{D}(b', \lambda)$ as b' approaches b. It generalizes a theorem of Hinton and Shaw [20, Lemma 2.1].

THEOREM 5.5. *Let $\mu_1(b') \leqq \cdots \leqq \mu_n(b')$ be the eigenvalues of $\mathcal{D}(b', \lambda)$. Let there be m, $n \leqq m \leqq 2n$ solutions of $JY' = (\lambda A + B)Y$, $\text{Im } \lambda \neq 0$, in $L_A^2(a, b)$. Then $\mu_1(b'), \cdots, \mu_{m-n}(b')$ remain finite and $\mu_{m-n+1}(b'), \cdots, \mu_{2n}(b')$ approach ∞ as b' approaches b.*

Proof. Suppose $\mu(b') < B$ for all b'. Let $v_{b'}$ be a unit eigenvector of $\mathcal{D}(b', \lambda)$. Setting $\chi_{b'} = \phi v_{b'}$, we get

$$2i \text{ Im } \lambda \int_a^{b'} \chi_{b'}^* A \chi_{b'} \, dx = v_{b'}^* \phi^* J \phi v_{b'} \Big|_a^{b'} = i \text{ sgn } (\text{Im } \lambda) \cdot \mu(b').$$

Thus

$$\int_a^{b'} \chi_{b'}^* A \chi_{b'} \, dx \leqq \mu(b')/|\text{Im } \lambda| \leqq B/|\text{Im } \lambda|.$$

Choosing a subsequence of $(v_{b'})$'s that converge, we find a solution $\chi = \phi v$ in $L_A^2(a, b)$. Since there are only mL_A^2-solutions and $\chi = \theta + \phi M$ comprises n of these, there can only be $m - n$ such χ's and only $m - n$ finite μ's.

Niessen [27] has a very similar theorem for proving the existence of square integrable solutions. His theorem depends upon the eigenvalues of

$$\begin{pmatrix} \mathscr{A} & \mathscr{B}^* \\ \mathscr{B} & \mathscr{D} \end{pmatrix}$$

instead of those of $\mathscr{D}$.

6. Singular boundary conditions. Unlike regular boundary points, such as b', where a regular boundary condition $(\beta_1, \beta_2)Y(b') = 0$ is imposed, singular boundary points, such as b, require a more careful approach. We define a singular boundary value that is valid on the domain of the maximal operator.

DEFINITION 6.1. We denote by D_M those elements Y in $L_A^2(a, b)$ satisfying

$$lY = JY' - BY = AF \tag{1}$$

exists almost everywhere for some F in $L_A^2(a, b)$.

DEFINITION 6.2. We define the maximal operator L_M by setting $L_M Y = F$ for all Y in D_M.

THEOREM 6.3. *Let Y_j be a solution of $JY_j' = (\bar{\lambda}A + B)Y_j$, $\operatorname{Im}\lambda \neq 0$. Then for all Y in D_M,*

$$B_{bj}(Y) = \lim_{x \to b} Y_j^* JY$$

exists if and only if Y_j is in $L_A^2(a, b)$.

Proof. Manipulation of

$$JY' - BY = AF$$

and

$$JY_j' - BY_j = \bar{\lambda} A Y_j$$

yields

$$(Y_j^* JY)' = Y_j^* A[F - \lambda Y].$$

Integrating, we get

$$(Y_j^* JY)(x) = Y_j^* JY(a) + \int_a^x Y_j^* A[F - \lambda Y]\, dx.$$

If Y_j is in $L_A^2(a, b)$, then as x approaches b, the integral on the right converges, and $\lim_{x \to b} Y_j^* JY(x)$ exists.

Conversely, if the integral on the right converges for all Y, F in $L_A^2(a, b)$, the Hahn-Banach theorem states that it generates a bounded operator. The Riesz Representation Theorem then affirms that Y_j is in $L_A^2(a, b)$.

DEFINITION 6.4. Let $\operatorname{Im}\lambda \neq 0$. Let $M(\bar{\lambda}) = \tilde{C}_0 + \tilde{R}_0 U R_0$ be on the limit circle. Let $\chi(x, \bar{\lambda}) = \theta(x, \bar{\lambda}) + \phi(x, \bar{\lambda})M(\bar{\lambda})$ satisfy (*) with λ replaced by $\bar{\lambda}$ and be in $L_A^2(a, b)$. We define the boundary value $B_\lambda(Y)$ by setting

$$B_\lambda(Y) = \lim_{x \to b} \chi(x, \bar{\lambda})^* JY(x)$$

for all Y in D_M.

Note that $B_\lambda(Y)$ is explicitly written as

$$B_\lambda(Y) = \lim_{x \to b} (I_n, M(\bar{\lambda})^*)\mathscr{Y}(x, \bar{\lambda})^* JY(x).$$

Note also that $\bar{\lambda}$ is used in the definition of $B_\lambda(Y)$. This is a convenience only. We shall remove the requirement later.

7. The differential operator L. We assume that the number of solutions of (*) in $L^2_A(a, b)$ in either half plane $\operatorname{Im}\lambda > 0$ and $\operatorname{Im}\lambda < 0$ is the same, m, as given by Theorems 5.4 and 5.5. This is certainly true if A and B are real. Using boundary conditions at both a and b, we can now define a self-adjoint differential operator on $L^2_A(a, b)$.

DEFINITION 7.1. We denote by D those elements Y in $L^2_A(a, b)$ satisfying

$$lY = JY' - BY = AF \tag{2}$$

exists almost everywhere for some F in $L^2_A(a, b)$,

$$(\alpha_1, \alpha_2)\, Y(a) = 0, \tag{3}$$

$$B_\lambda(Y) = 0. \tag{4}$$

DEFINITION 7.2. We define the operator L by setting $LY = F$ for all Y in D.

The inverse of $L - \lambda I$ can now be calculated easily. If we solve

$$JY' = (\lambda A + B)\, Y + AF$$

by variation of parameters, the substitution of $Y = \mathcal{Y}C$, where $\mathcal{Y}$ is the fundamental matrix, yields $J\mathcal{Y}C' = AF$, so $C' = -J\mathcal{Y}^t AF$ where we have written $\mathcal{Y}^t(x, \lambda)$ for $\mathcal{Y}(x, \bar{\lambda})^*$, since if the matrix coefficients A and B are real, $\mathcal{Y}^t = \mathcal{Y}^T$, the transpose. Integrating, we find

$$Y = -\mathcal{Y}(x, \lambda) \int_a^x J\mathcal{Y}^t(\xi, \lambda) A(\xi x) F(\xi)\, d\xi + \mathcal{Y}(x, \lambda) K$$

where K is constant.

Since

$$\begin{pmatrix} \alpha_1 & \alpha_2 \\ 0 & 0 \end{pmatrix} Y(a) = 0,$$

$$0 + \begin{pmatrix} \alpha_1 & \alpha_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1^* & -\alpha_2^* \\ \alpha_2^* & \alpha_1^* \end{pmatrix} K = 0,$$

or

$$\begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} K = 0.$$

Further,

$$\begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(x, \lambda) JY(x) = -\begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(x, \lambda) J\mathcal{Y}(x, \lambda) \int_a^x J\mathcal{Y}^t(\xi, \lambda) A(\xi) F(\xi)\, d\xi$$
$$+ \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(x, \lambda) J\mathcal{Y}(x, \lambda) K.$$

Since $\mathcal{Y}^t(x, \lambda) J\mathcal{Y}(x, \lambda) \equiv J$ for all x,

$$\begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(x, \lambda) J\mathcal{Y}(x) = \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \int_a^x \mathcal{Y}^t(\xi, \lambda) A(t) F(t)\, dt$$
$$+ \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} K.$$

Letting $x \to b$,

$$0 = \int_a^b \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(\xi, \lambda) A(\xi) F(\xi)\, d\xi + \begin{pmatrix} 0 & 0 \\ M^t & -I_n \end{pmatrix} K.$$

Adding

$$0 = \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} K$$

to this, we have

$$0 = \int_a^b \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(\xi, \lambda) A(\xi) F(\xi)\, d\xi + \begin{pmatrix} I_n & 0 \\ M^t & -I_n \end{pmatrix} K.$$

The coefficient of K is its own inverse. Hence

$$K = \int_a^b \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(\xi, \lambda) A(\xi) F(\xi)\, d\xi,$$

and

$$Y = \mathcal{Y}(x, \lambda) \int_a^x \begin{pmatrix} 0 & I_n \\ 0 & M^t \end{pmatrix} \mathcal{Y}^t(\xi, \lambda) A(\xi) F(\xi)\, d\xi + \mathcal{Y}(x, \lambda) \int_x^b \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(\xi, \lambda) A(\xi) F(\xi)\, d\xi.$$

At this point we note that the Green's function, the kernel of the integral operator, is the limit of Green's functions of regular Sturm–Liouville problems. Since regular problems satisfy

$$G(\lambda, x, \xi) = G^*(\bar{\lambda}, \xi, x),$$

so does the present Green's function. Comparison of the two sides gives a definition of $M(\lambda)$ in terms of $M(\bar{\lambda})$,

$$M(\lambda) = M^t = M^*(\bar{\lambda}),$$

and $\chi(x, \lambda) = \theta(x, \lambda) + \phi(x, \lambda) M(\lambda)$ is in $L_A^2(a, b)$. Now from the first integral, the terms

$$\mathcal{Y}(x, \lambda) \begin{pmatrix} 0 & I_n \\ 0 & M^t \end{pmatrix} \mathcal{Y}^*(\xi, \lambda) = \chi(x, \lambda) \phi^t(\xi, \lambda).$$

From the second, recalling that $M^t = M$,

$$\mathcal{Y}(x, \lambda) \begin{pmatrix} 0 & 0 \\ I_n & M^t \end{pmatrix} \mathcal{Y}^t(\xi, \lambda) = \phi(x, \lambda) \chi^t(\xi, \lambda).$$

Therefore

$$Y(x) = \chi(x, \lambda) \int_a^x \phi(\xi, \bar{\lambda})^* A(\xi) F(\xi)\, d\xi + \phi(x, \lambda) \int_x^b \chi(\xi, \bar{\lambda})^* A(\xi) F(\xi)\, d\xi,$$

or

$$Y = \int_a^b G(\lambda, x, \xi) A(\xi) F(\xi)\, d\xi,$$

where

$$(**)\qquad \begin{aligned} G(\lambda, x, \xi) &= \chi(x, \lambda)\phi(\xi, \bar{\lambda})^*, \qquad a \leqq \xi \leqq x \leqq b \\ &= \phi(x, \lambda)\chi(\xi, \bar{\lambda})^*, \qquad a \leqq x \leqq \xi \leqq b. \end{aligned}$$

THEOREM 7.3. *The parameter λ, used in defining the boundary condition B_λ, is in the resolvent of L,*

$$(L-\lambda I)^{-1}(F(x)) = \int_0^b G(\lambda, x, \xi)A(\xi)F(\xi)\, d\xi,$$

where G is given by (**).

THEOREM 7.4. *The operator L is self-adjoint.*

Proof. Let $(L-\lambda I)Y = F$, $(L^*-\bar{\lambda}I)Z = G$. Then

$$\begin{aligned} \langle (L-\lambda I)^{-1}F, G\rangle &= \int_a^b G^*(x)A(x) \int_a^b G(\lambda, x, \xi)A(\xi)F(\xi)\, d\xi\, dx \\ &= \int_a^b \left[\int_a^b G^*(\lambda, x, \xi)A(x)G(x)\, dx\right]^* A(\xi)F(\xi)\, d\xi \\ &= \int_a^b [G^*(\lambda, \xi, x)A(\xi)G(\xi)\, d\xi]^* A(x)F(x)\, dx \\ &= \int_a^b \left[\int_a^b G(\bar{\lambda}, x, \xi)A(\xi)G(\xi)\, d\xi\right]^* A(x)F(x)\, dx \\ &= \langle F, (L-\bar{\lambda}I)^{-1}G\rangle \end{aligned}$$

since $G(\lambda, x, \xi) = G(\bar{\lambda}, \xi, x)^*$. But

$$\langle (L-\lambda)^{-1}F, G\rangle = \langle F, (L^*-\bar{\lambda}I)G\rangle,$$

so $(L-\bar{\lambda}I)^{-1} = (L^*-\bar{\lambda}I)^{-1}$. Taking inverses, $L-\bar{\lambda}I = L^*-\bar{\lambda}I$. Canceling $\bar{\lambda}I$, $L = L^*$.

We remark that when $M(\bar{\lambda})$ is on the limit circle, then $\lim_{x\to b}\chi(x, \bar{\lambda})^*J\chi(x, \bar{\lambda}) = 0$. This is a test for self-adjointness of the singular boundary condition $B_\lambda(Y)$. It is the limiting form of $(\beta_1, \beta_2)J(\beta_1, \beta_2)^* = 0$, required in the regular case.

THEOREM 7.5. *The operator $(L-\lambda I)^{-1}$ is a bounded operator.*

$$\|(L-\lambda I)^{-1}\| \leqq 1/|\operatorname{Im}\lambda|.$$

Proof. Let $(L-\lambda I)Y = F$. Then

$$\langle Y, F\rangle - \langle F, Y\rangle = \langle Y, (L-\lambda I)Y\rangle - \langle (L-\lambda I)Y, Y\rangle = (\lambda-\bar{\lambda})\langle Y, Y\rangle.$$

Applying Schwarz's inequality to the left,

$$2|\operatorname{Im}\lambda|\, \|Y\|^2 \leqq 2\|Y\|\, \|F\|.$$

Canceling $\|Y\|$ yields the result. (This may be found in [1] as well.)

THEOREM 7.6. *If $JY' - BY = AF$, $AY = 0$ implies $Y = 0$, then D is dense in $L^2_A(a, b)$.*

Proof. If D is not dense, then there is a G orthogonal to D. Let Y be in D, and let Z satisfy $Z \in D$, $JZ' - BZ = \bar{\lambda}AZ + AG$ for $\operatorname{Im}\lambda \neq 0$. Then

$$\begin{aligned} 0 = \langle Y, G\rangle &= \int_a^b G^*AY\, dx = \int_a^b [JZ' - BZ - \bar{\lambda}AZ]^* Y\, dx \\ &= \int_a^b Z^*[JY' - BY - \lambda AY]\, dx. \end{aligned}$$

Let $JY' - BY - \lambda AY = AF$. Then

$$0 = \langle F, Z \rangle = \int_a^b Z^* AF \, dx.$$

Now F is arbitrary, so let $F = Z$. Thus $\int_a^b Z^* AZ \, dx = 0$ and $AZ = 0$. Thus $JZ' - BZ = AG$, and $Z = 0$. Since $Z = 0$, $AG = 0$, and $G = 0$ in $L_A^2(a, b)$.

8. Extension of the boundary conditions. We have chosen $\bar{\lambda}$ to be fixed in generating the boundary condition at $x = b$. We would like to show now that, properly extended, $\chi(x, \lambda) = \theta(x, \lambda) + \phi(x, \lambda)M(\lambda)$ remains in $L_A^2(a, b)$ and that if $\lim_{x \to b} \chi(x, \bar{\lambda})^* JY(x) = 0$, then for all λ, $\operatorname{Im} \lambda \neq 0$, $\lim_{x \to b} \chi(x, \bar{\lambda})^* JY(x) = 0$. In a sense, the boundary condition is independent of λ.

In the course of deriving the Green's function, we show that $M(\lambda) = M(\bar{\lambda})^*$. As a consequence we find the following theorem holds.

THEOREM 8.1. *If* $\chi(x, \lambda) = \theta(x, \lambda) + \phi(x, \lambda)M(\lambda)$, *then* $\lim_{x \to b} \chi(x, \bar{\lambda})^* \cdot J\chi(x, \lambda) = 0$.

Proof.

$$\begin{aligned} \chi(x, \bar{\lambda})^* J \chi(x, \lambda) &= (I_n, M(\bar{\lambda})^*) \mathcal{Y}(x, \bar{\lambda})^* J \mathcal{Y}(x, \lambda) \begin{pmatrix} I_n \\ M(\lambda) \end{pmatrix} \\ &= (I_n, M(\bar{\lambda})^*) J \begin{pmatrix} I_n \\ M(\lambda) \end{pmatrix} \\ &= 0, \end{aligned}$$

so the limit is zero also.

COROLLARY 8.2. *Let the columns of* $\chi(x, \lambda)$ *be modified smoothly so that they vanish near* $x = a$. *Then, so modified, the columns of* $\chi(x, \lambda)$ *are in* D.

THEOREM 8.3. *If* Y *is in* D, *then* $\lim_{x \to b} \chi(x, \lambda)^* JY(x) = 0$.

Proof. Since each column of $\chi(x, \lambda)$, appropriately modified, is in D, an application of Green's formula shows the limit is zero.

DEFINITION 8.4. We extend the definition of $\chi(x, \bar{\lambda})$ to other values $\bar{\mu}$ in the same half plane as $\bar{\lambda}$, by

$$\chi(x, \bar{\mu}) = \chi(x, \bar{\lambda}) + (\bar{\mu} - \bar{\lambda}) \int_a^b G(\bar{\lambda}, x, \xi) A(\xi) \chi(\xi, \mu) \, d\xi.$$

It is well known that if $|\bar{\mu} - \bar{\lambda}| < |1/\operatorname{Im} \lambda|$, then with initial estimate $\chi(x, \bar{\lambda})$, a Neumann series may be used to show that the integral equation has a unique solution, analytic in $\bar{\mu}$ and in $L_A^2(a, b)$. Further, this solution can be extended analytically through the entire half plane containing $\bar{\lambda}$. It is also easy to show that $(L - \bar{\mu})\chi(x, \bar{\mu}) = 0$.

A simple calculation shows that $\chi(x, \bar{\mu}) = \theta(x, \bar{\mu}) + \phi(x, \bar{\mu})M(\bar{\mu})$, where

$$M(\bar{\mu}) = M(\bar{\lambda}) + (\bar{\mu} - \bar{\lambda}) \int_a^b \chi(x, \bar{\lambda})^* A(\xi) \chi(\xi, \bar{\mu}) \, d\xi,$$

thus extending $M(\bar{\mu})$ as well.

THEOREM 8.5. *For different parameters* $\bar{\lambda}$ *and* $\bar{\mu}$, $\lim_{x \to b} \chi(x, \bar{\lambda})^* J \chi(x, \bar{\mu}) = 0$.

Proof. If $\lim_{x \to b} \chi(x, \bar{\lambda})^* J$ is applied to the integral equation, both terms on the right-hand side have limit zero. Extended analytically in $\bar{\mu}$, this remains zero.

COROLLARY 8.6. *Let the columns of* $\chi(x, \bar{\mu})$ *be modified smoothly so that they vanish near* $x = a$. *Then, so modified, the columns of* $\chi(x, \bar{\mu})$ *are in* D.

THEOREM 8.7. *If* Y *is in* D, *then* $\lim_{x \to b} \chi(x, \bar{\mu})^* JY(x) = 0$.

DEFINITION 8.8. We extend the definition of $\chi(x, \mu)$ to other values μ in the same half plane as λ by

$$\chi(x, \mu) = \chi(x, \lambda) + (\mu - \lambda) \int_a^b G(\lambda, x, \xi) A(\xi) \chi(\xi, \mu)\, d\xi.$$

The comments made after Definition 8.4 apply here as well with the conjugate signs removed.

THEOREM 8.9. *For different parameters λ and μ, $\lim_{x \to b} \chi(x, \lambda)^* J \chi(x, \mu) = 0$.*

COROLLARY 8.10. *Let the columns of $\chi(x, \mu)$ be modified smoothly so that they vanish near $x = a$. Then, so modified, the columns of $\chi(x, \mu)$ are in D.*

Proof. Clearly the columns are in D, associated with λ. But D, defined by $B_\lambda(Y)$, and D, defined by $B_{\bar\lambda}(Y)$, are the same.

THEOREM 8.11. *If Y is in D, then $\lim_{x \to b} \chi(x, \mu) J Y(x) = 0$.*

In summary, we state the following theorem.

THEOREM 8.12. *Let Y be in D. Then, for all λ, $\operatorname{Im} \lambda \neq 0$, $B_\lambda(Y) = 0$.*

THEOREM 8.13. *For all λ, $\operatorname{Im} \lambda \neq 0$, $(L - \lambda I)^{-1}$ exists and is given by (**).*

$$\|(L - \lambda I)^{-1}\| \leqq 1/|\operatorname{Im} \lambda|.$$

Proof. The previous calculation holds no matter what λ.

Finally we comment that for all μ, $\operatorname{Im} \mu \neq 0$, $\chi(x, \mu) = \theta(x, \mu) + \phi(x, \mu) M(\mu)$ satisfies $\lim_{x \to b} \chi(x, \mu)^* J \chi(x, \mu) = 0$. This implies that $M(\mu)$ is on the μ limit circle; however, its exact relation to the sequence (β_1, β_2) and U, used to define $M(\bar\lambda)$ on the original limit circle, is not clear.

9. The Lagrange bilinear form. Just as L is a restriction of the maximal operator L_M, all other self-adjoint operators involving the differential equation (*) are also generated by examining other restrictions of L_M. This can be done by examining Green's formula, a major component of which is the Lagrange bilinear form $Z^* J Y$. This must be expressed properly in order to identify those which are self-adjoint. If Y and Z are in D_M, then a preliminary form of Green's formula is

$$\int_a^b [Z^*(JY' - BY) - (JZ' - BZ)^* Y]\, dx = Z^* J Y|_a^b.$$

Since Y and Z are in D_M, they satisfy

$$JY' - (\lambda A + B) Y = AF, \qquad JZ' - (\bar\lambda A + B) Z = AG,$$

where $\operatorname{Im} \lambda > 0$ and F and G are in $L_A^2(a, b)$. Using the Green's function for L we can write

$$Y(x) = \int_a^b G(\lambda, x, \xi) A(\xi) F(\xi)\, d\xi + \chi_b(x, \lambda) C_1 + \phi(x, \lambda) C_2,$$

$$Z(x) = \int_a^b G(\bar\lambda, x, \xi) A(\xi) G(\xi)\, d\xi + \chi_b(x, \bar\lambda) D_1 + \phi(x, \bar\lambda) D_2,$$

where C_2 and D_2 are chosen so that $\phi(x, \lambda) C_2$ and $\phi(x, \bar\lambda) D_2$ consists only of $L_A^2(a, b)$ solutions.

We can solve for C_1, C_2, D_1, D_2. Since the integral in the formula for Y, as well as the term $\chi_b(x, \lambda) C_1$ satisfy the limit boundary condition, and $\lim_{x \to b} \chi_b(x, \bar\lambda) J \phi(x, \lambda) = -I$,

$$\lim_{x \to b} \chi_b(x, \bar\lambda)^* J Y(x) = -C_2.$$

Further

$$\lim_{x\to b} D_2^*\phi(x,\bar\lambda)^*JY = \lim_{x\to b} D_2^*\phi(x,\bar\lambda)^*JR_Y(x) + D_2^*C_1,$$

where

$$R_Y(x) = \int_a^b G(\lambda, x, \xi)A(\xi)F(\xi)\,d\xi.$$

Likewise

$$\lim_{x\to b} Z(x)^*J\chi_b(x,\lambda) = D_2^*$$

and

$$\lim_{x\to b} Z(x)^*J\phi(x,\lambda)C_2 = \lim_{x\to b} R_Z(x)^*J\phi(x,\lambda)C_2 - D_1^*C_2,$$

where

$$R_Z(x) = \int_a^b G(\bar\lambda, x, \xi)A(\xi)G(\xi)\,d\xi.$$

We then compute

$$\lim_{x\to b} Z^*(x)JY(x) = \lim_{x\to b} R_Z(x)^*J\phi(x,\lambda)C_2 + \lim_{x\to b} D_2^*\phi(x,\bar\lambda)^*JR_Y(x) - D_1^*C_2 + D_2^*C_1,$$

the other terms cancelling because they satisfy the χ boundary condition at b.

Eliminating the terms $D_1^*C_2$ and $D_2^*C_1$ by substitution, we obtain

$$\lim_{x\to b} Z(x)^*JY(x) = \lim_{x\to b} Z(x)^*J\phi(x,\lambda)C_2 + \lim_{x\to b} D_2^*\phi(x,\bar\lambda)^*JY(x).$$

Now the term

$$\phi(x,\lambda)C_2 = -\phi(x,\lambda)\left[\lim_{x\to b}\chi_b(x,\bar\lambda)^*JY(x)\right].$$

Let $\phi(x,\lambda) = (\phi_1, \phi_2)(x)E_b$, where $\phi_1(x,\lambda)$ consists of those $m-n$ solutions of $(*)$ in $L_A^2(a,b)$, and $\phi_2(x,\lambda)$ are those not in $L_A^2(a,b)$. Let

$$\chi_b(x,\bar\lambda)E_b^* = (\chi_1, \chi_2)(x,\bar\lambda),$$

where χ_1 contains $m-n$ components as well. Then

$$\phi(x,\lambda)C_2 = -\phi_1(x,\lambda)\left[\lim_{x\to b}\chi_1(x,\bar\lambda)^*JY(x)\right].$$

Likewise, the term

$$D_2^*\phi(x,\bar\lambda)^* = \left[\lim_{x\to b} Z(x)^*J\chi_b(x,\lambda)\right]\phi(x,\bar\lambda)^*$$

$$= \left[\lim_{x\to b} Z(x)^*J\chi_1(x,\lambda)\right]\phi_1(x,\bar\lambda)^*$$

where $\phi_1(x,\bar\lambda)$ contains $m-n$ components. The terms

$$-\phi_2(x,\phi)\left[\lim_{x\to b}\chi_2(x,\bar\lambda)^*JY(x)\right]$$

and

$$\left[\lim_{x\to b} Z(x)^*J\chi_2(x,\lambda)\right]\phi_2(x,\bar\lambda)^*$$

are not present because $\phi(x,\lambda)C_2$ and $\phi(x,\bar{\lambda})D_2$ contain only solutions in $L^2_A(a,b)$. Therefore, we have the following theorem.

THEOREM 9.1. *Let* $\phi(x,\lambda)=(\phi_1,\phi_2)(x,\lambda)E_b$, *where* ϕ_1 *consists of all of the* $\phi - L^2_A(a,b)$ *solutions of* $(*)$. *Let* $\phi(x,\bar{\lambda})=(\phi_1,\phi_2)(x,\bar{\lambda})\tilde{E}_b$, *where* ϕ_1 *consists of all of the* $\phi - L^2_A(a,b)$ *solutions of* $(*)$ *with* λ *replaced by* $\bar{\lambda}$. *Let* $(\chi_1,\chi_2)(x,\bar{\lambda})=\chi_b(x,\bar{\lambda})E^*_b$, *and let* $(\chi_1,\chi_2)(x,\lambda)=\chi_b(x,\lambda)\tilde{E}^*_b$. *Then for all* Y *and* Z *in* D_M

$$\lim_{x\to b}\chi_2(x,\bar{\lambda})^*JY(x)=0, \qquad \lim_{x\to b}\chi_2(x,\lambda)^*JZ(x)=0.$$

Proof. This is the only way to ensure that $\phi_2(x,\lambda)$ and $\phi_2(x,\bar{\lambda})$ are not present.

The possibility of this occurring was reported to the author by D. B. Hinton several years ago.

Theorem 9.1 explains an apparent discrepancy. It appears that by imposing the conditions $\lim_{x\to b}\chi(x,\bar{\lambda})^*JY(x)=0$ and $(\alpha_1,\alpha_2)Y(a)=0$, that $2n$ boundary conditions are being required. However, since the deficiency indices at $x=b$ are (m,m), only $n+m$ should be imposed.

Theorem 9.1 shows that in fact when $m-n$ ϕ combinations are in $L^2_A(a,b)$, $2n-m$ combinations are not, and so there are $2n-m$ χ_2 automatically zero conditions. As a result χ imposes only $n-(2n-m)=m$ constraints. With n constraints at $x=a$, the total is $n+m$, not $2n$.

Returning to $\lim_{x\to b}Z(x)^*JY(x)$, we find

$$\lim_{x\to b}Z(x)^*JY(x)=\left[\lim_{x\to b}\phi_1(x,\lambda)^*JZ(x)\right]^*\left[\lim_{x\to b}\chi_1(x,\bar{\lambda})^*JY(x)\right]$$
$$-\left[\lim_{x\to b}\chi_1(x,\lambda)^*JZ(x)\right]^*\left[\lim_{x\to b}\phi_1(x,\bar{\lambda})^*JY(x)\right].$$

THEOREM 9.2. *Under the conditions of Theorem* 9.1, *let*

$$B_b(Y)=\begin{pmatrix}\lim_{x\to b}\chi_1(x,\bar{\lambda})^*JY(x)\\ \lim_{x\to b}\phi_1(x,\bar{\lambda})^*JY(x)\end{pmatrix},$$

$$\tilde{B}_b(Z)=\begin{pmatrix}\lim_{x\to b}\chi_1(x,\lambda)^*JZ(x)\\ \lim_{x\to b}\phi_1(x,\lambda)^*JZ(x)\end{pmatrix},$$

$$J_b=\begin{pmatrix}0 & -I_{m-n}\\ I_{m-n} & 0\end{pmatrix},$$

where (m,m) *are the defect indices of* $(*)$ *at* b; *then*

$$\lim_{x\to b}Z(x)^*JY(x)=\tilde{B}_b(Z)^*J_bB_b(Y).$$

The symbol $\tilde{}$ indicates λ is used instead of $\bar{\lambda}$.

If the end $x=a$ is regular, the computation is a bit easier. Note that

$$[\chi_b(x,\lambda),\phi(x,\lambda)]J[\chi_b(x,\bar{\lambda}),\phi(x,\bar{\lambda})]^*=J.$$

Therefore

$$Z^*(a)JY(a)=-Z^*(a)J[\chi_b(a,\lambda),\phi(a,\lambda)]J[\chi_b(a,\bar{\lambda}),\phi(a,\bar{\lambda})]^*JY(a).$$

THEOREM 9.3. *Let*

$$B_a(Y)=\begin{pmatrix}\chi_b(a,\bar{\lambda})^*JY(a)\\ \phi_b(a,\bar{\lambda})^*JY(a)\end{pmatrix},$$

$$\tilde{B}_a(Z)=\begin{pmatrix}\chi_b(a,\lambda)^*JZ(a)\\ \phi_b(a,\lambda)^*JZ(a)\end{pmatrix},$$

$$J=\begin{pmatrix}0 & -I_n\\ I_n & 0\end{pmatrix};$$

then

$$Z^*(a)JY(a)=\tilde{B}_a(Z)^*JB_a(Y).$$

The terms $\phi_b(a,\bar{\lambda})JY(a)$ and $\phi_b(a,\lambda)JZ(a)$ are simply $(\alpha_1,\alpha_2)Y(a)$ and $(\alpha_1,\alpha_2)Z(a)$. The terms $\chi_b(a,\bar{\lambda})^*JY(a)$ and $\chi_b(a,\lambda)^*JZ(a)$ may be replaced by $\theta(a,\bar{\lambda})^*JY(a)$ and $\theta(a,\lambda)^*JZ(a)$, and the formula for $Z(a)^*JY(a)$ still holds. $\theta(a,\bar{\lambda})^*JY(a)=(-\alpha_2,\alpha_1)Y(a)$; $\theta(a,\bar{\lambda})^*JZ(a)=(-\alpha_2,\alpha_1)Z(a)$ as well. These are complimentary boundary forms.

10. Green's formula. If the pieces from the previous section are assembled together, we have the following theorem.

THEOREM 10.1. *Let Y and Z be in D_M. Then*

$$\int_a^b\{Z^*[JY'-BY]-[JZ'-BZ]^*Y\}\,dx=-\tilde{B}_a(Z)^*JB_a(Y)+\tilde{B}_b(Z)^*J_bB_b(Y)$$

$$=(\tilde{B}_a(Z)^*,\tilde{B}_b(Z)^*)\begin{pmatrix}-J & 0\\ 0 & J_b\end{pmatrix}\begin{pmatrix}B_a(Y)\\ B_b(Y)\end{pmatrix}.$$

Let M and N be $r\times 2n$ and $r\times(2m-2n)$ matrices, respectively, where (m,m) are the defect indices of $(*)$ at b, with rank $(MN)=r$, $0\leqq r\leqq 2m$. Further, let P and Q be $(2m-r)\times 2n$ and $(2m-r)\times(2m-2n)$ matrices such that

$$\begin{pmatrix}M & N\\ P & Q\end{pmatrix}$$

is nonsingular. If $\tilde{M},\tilde{N},\tilde{P},\tilde{Q}$ are $r\times 2n$, $r\times(2m-2n)$, $(2m-r)\times 2n$, $(2m-r)\times(2m-2n)$ matrices such that

$$\begin{pmatrix}\tilde{M}^* & \tilde{P}^*\\ \tilde{N}^* & \tilde{Q}^*\end{pmatrix}\begin{pmatrix}M & N\\ P & Q\end{pmatrix}=\begin{pmatrix}-J & 0\\ 0 & J_b\end{pmatrix},$$

then inserting this in Green's formula gives it its final form.

THEOREM 10.2. (Green's formula.) *Let Y and Z be in D_M. Then*

$$\int_a^b\{Z^*[JY'-BY]-[JZ'-BZ]^*Y\}\,dx$$

$$=[\tilde{M}\tilde{B}_a(Z)+\tilde{N}\tilde{B}_b(Z)]^*[MB_a(Y)+NB_b(Y)]$$

$$+[\tilde{P}\tilde{B}_a(Z)+\tilde{Q}\tilde{B}_b(Z)]^*[PB_a(Y)+QB_a(Y)].$$

DEFINITION 10.3. We denote by $\tilde{D}$ those elements Y in $L_A^2(a, b)$ satisfying
(1) Y is in D_M;
(2) $MB_a(Y)+NB_b(Y)=0$.

DEFINITION 10.4. We denote by $\tilde{L}$ the operator defined by setting $\tilde{L}Y=F$ whenever $JY'-BY=AF$ and Y is in $\tilde{D}$.

DEFINITION 10.5. We denote by $\tilde{D}^*$ those elements Z in $L_A^2(a, b)$ satisfying
(1) Z is in D_M;
(2) $\tilde{P}\tilde{B}_a(Z)+\tilde{Q}\tilde{B}_b(Z)=0$.

DEFINITION 10.6. We denote by $\tilde{L}^*$ the operator defined by setting $\tilde{L}^*Z=G$ whenever $JZ'-BZ=AG$ and Z is in $\tilde{D}^*$.

THEOREM 10.7. *The abuse of notation above is correct. The adjoint of $\tilde{L}$ in $L_A^2(a, b)$ is $\tilde{L}^*$. The adjoint of $\tilde{L}^*$ in $L_A^2(a, b)$ is $\tilde{L}$.*

Proof. The form of the adjoint of $\tilde{L}$ is well known to be the same as that of $\tilde{L}^*$. From Green's formula it is clear that $\tilde{D}^*$ is included in domain of the adjoint.

Conversely, again from Green's formula, since $PB_a(Y)+QB_b(Y)$ is arbitrary, any element in the adjoint's domain must be in $\tilde{D}^*$.

There are parametric boundary conditions as well. Set

$$MB_a(Y)+NB_b(Y)=0, \qquad PB_a(Y)+QB_b(Y)=\Gamma,$$

where Γ is arbitrary. Multiply

$$\begin{pmatrix} M & N \\ P & Q \end{pmatrix}\begin{pmatrix} B_a(Y) \\ B_b(Y) \end{pmatrix}=\begin{pmatrix} 0 \\ \Gamma \end{pmatrix}$$

by

$$\begin{pmatrix} J & 0 \\ 0 & -J_{m-n} \end{pmatrix}\begin{pmatrix} \tilde{M}^* & \tilde{P}^* \\ \tilde{N}^* & \tilde{Q}^* \end{pmatrix}.$$

The result is

$$B_a(Y)=J\tilde{P}^*\Gamma, \qquad B_b(Y)=-J_b\tilde{Q}^*\Gamma.$$

Likewise, if

$$\tilde{M}\tilde{B}_a(Z)+\tilde{N}\tilde{B}_b(Z)=\Delta, \qquad \tilde{P}\tilde{B}_a(Z)+\tilde{Q}\tilde{B}_b(Z)=0,$$

then post-multiply

$$(\tilde{B}_a(Z)^*, \tilde{B}_b^*(Z))\begin{pmatrix} \tilde{M}^* & \tilde{P}^* \\ \tilde{N}^* & \tilde{Q}^* \end{pmatrix}=(\Delta^*, 0)$$

by

$$\begin{pmatrix} M & N \\ P & Q \end{pmatrix}\begin{pmatrix} J & 0 \\ 0 & -J_b \end{pmatrix}.$$

The result is

$$\tilde{B}_a(Z)=-JM^*\Delta, \qquad \tilde{B}_b(Z)=J_bN^*\Delta.$$

THEOREM 10.8. *The parametric boundary conditions are fully equivalent to the original boundary conditions.*

11. Equivalent boundary conditions. The evaluation of $\lim_{x\to b} Z(x)^*JY(x)$, found in the proof of Theorem 9.1, enables us to establish a number of interesting relationships.

First think of it as defining a boundary condition on Y:

$$B_Z(Y) = \lim_{x\to b} Z(x)^* JY(x)$$

$$= \left[\lim_{x\to b} \phi_1(x,\lambda)^* JZ(x)\right]^* \left[\lim_{x\to b} \chi_1(x,\bar{\lambda})^* JY(x)\right]$$

$$- \left[\lim_{x\to b} \chi_1(x,\lambda) JZ(x)\right]^* \left[\lim_{x\to b} \phi_1(x,\bar{\lambda}) JY(x)\right].$$

If we let

$$C = \left[\left\{\lim_{x\to b} \phi_1(x,\lambda) JZ(x)\right\}^*, \left\{-\lim_{x\to b} \chi_1(x,\lambda)^* JZ(x)\right\}^*\right]$$

we have the following theorem.

THEOREM 11.1. *Let Z be in D_M; then*

$$B_Z(Y) = CB_b(Y).$$

Since Glazman showed that all self-adjoint boundary conditions can be generated by using appropriate elements from D_M, this shows that our solution-generated boundary conditions also suffice.

As a special case, let $Z = \chi_1(x,\lambda)$.

THEOREM 11.2. *Let $\tilde{V}_{11} = [\lim_{x\to b} \phi_1(x,\lambda)^* J\chi_1(x,\lambda)]^*$; then*

$$\left[\lim_{x\to b} \chi_1(x,\lambda)^* JY(x)\right] = \tilde{V}_{11}\left[\lim_{x\to b} \chi_1(x,\bar{\lambda})^* JY(x)\right].$$

The coefficient

$$\tilde{V}_{12} = \left[\lim_{x\to b} \chi_1(x,\lambda)^* J\chi_1(x,\lambda)]^*\right]$$

is zero.

Likewise, let $Z = \phi_1(x,\lambda)$.

THEOREM 11.3. *Let*

$$\tilde{V}_{21} = \left[\lim_{x\to b} \phi_1(x,\lambda)^* J\phi_1(x,\lambda)\right]^*,$$

$$\tilde{V}_{22} = \left[\lim_{x\to b} \chi_1(x,\lambda)^* J\phi_1(x,\lambda)\right]^*,$$

and $\tilde{V} = (\tilde{V}_{ij})$; then

$$\tilde{B}_b(Y) = \tilde{V}B_b(Y).$$

COROLLARY 11.4. *There exists a matrix V such that*

$$B_b(Y) = V\tilde{B}_b(Y).$$

Note that if λ can be chosen to be real, which is frequently the case, then $V = \tilde{V} = I$.

12. Self-adjointness.

THEOREM 12.1. *$\tilde{L}$ is self-adjoint if and only if $r = m$ and*

$$MJM^* = NVJ_bN^*.$$

Proof. Suppose $\tilde{L}$ is self-adjoint. Then Y in $\tilde{D}$ satisfies

$$MB_a(Y)+NB_b(Y)=0$$

and

$$\tilde{P}\tilde{B}_a(Y)+\tilde{Q}\tilde{B}_b(Y)=0.$$

Since the boundary evaluations at each end must be the same, and $B_a(Y)=\tilde{B}_a(Y)$, we find $B_b(Y)=V\tilde{B}_b(Y)$ for all Y in $\tilde{D}$.

Then

$$\tilde{B}_a(Y)=B_a(Y)=-JM^*\Delta, \qquad \tilde{B}_b(Y)=V^{-1}B_b(Y)=J_bN^*\Delta.$$

We have

$$M(-JM^*\Delta)+N(VJ_bN^*\Delta)=0, \qquad \text{or} \qquad [-MJM^*+NVJ_bN^*]\Delta=0.$$

Since Δ may be arbitrary,

$$MJM^*=NVJ_bN^*.$$

Conversely, if

$$MJM^*=NVJ_bN^*,$$

then

$$(M,N)\begin{pmatrix}-JM\\VJ_bN^*\end{pmatrix}=0, \qquad (M,N)\begin{pmatrix}B_a(Y)\\B_b(Y)\end{pmatrix}=0.$$

Thus there must exist a nonsingular Δ such that

$$\begin{pmatrix}-JM^*\\VJ_bN^*\end{pmatrix}\Delta=\begin{pmatrix}B_a(Y)\\B_b(Y)\end{pmatrix}$$

and $B_a(Y)=-JM^*\Delta$, $B_b(Y)=VJ_bN^*\Delta$. Hence $\tilde{B}_a(Y)=-JM^*\Delta$, $\tilde{B}_b(Y)=J_bN^*\Delta$. This implies

$$\tilde{P}\tilde{B}_a(Y)+\tilde{Q}\tilde{B}_b(Y)=0,$$

and so Y is in $\tilde{D}^*$. A symmetric argument then shows $\tilde{D}=\tilde{D}^*$, and $\tilde{L}$ is self-adjoint.

13. The spectral resolution of L. This section closely follows the lead of Coddington and Levinson [3, Chap. 9].

If we consider the regular boundary value problem on $[a, b']$,

$$JY'(\lambda A+B)Y,$$

$$(\alpha_1\alpha_2)Y(a)=0, \qquad (\beta_1\beta_2)Y(b')=0,$$

we recall that ϕ satisfies the boundary condition at a, while for $\operatorname{Im}\lambda\neq 0$, $\chi_{b'}=\theta+\phi M_{b'}$ satisfies the boundary condition at b provided

$$M_{b'}=-[(\beta_1\beta_2)\phi(b',b)]^{-1}[(\beta_1\beta_2)\theta(b',\lambda)].$$

For all λ, $M_{b'}$ is analytic in λ except for real poles, which are eigenvalues for the regular problem on $[a, b']$. We denote these eigenvalues by $\{\lambda_k\}_{k=1}^{\infty}$. Their corresponding eigenfunctions are $\{V_k\}_{k=1}^{\infty}$. For each λ_k,

$$V_k=\phi(x,\lambda_k)K_k,$$

where K_k is an $n \times 1$ matrix. We assume that at multiple eigenvalues, the corresponding eigenfunctions have been made orthogonal.

THEOREM 13.1. *Let $F(x)$ be an arbitrary element in $L_A^2(a, b')$. Then*

$$F(x) = \sum_{k=1}^{\infty} \phi_k(x) C_k,$$

where $\phi_k(x) = \phi(x, \lambda_k)$ and

$$C_k = \left[\int_a^{b'} \phi_k^* A \phi_k \, d\xi \right]^{-1} \int_a^{b'} \phi_k A F \, d\xi.$$

This is, of course, just the standard eigenfunction expansion. Note that C_k is an $n \times 1$ matrix.

THEOREM 13.2. *Let $F(x)$ be an arbitrary element in $L_A^2(a, b')$. Then*

$$\int_a^{b'} F^* A F \, d\xi = \sum_{k=1}^{\infty} \left[\int_a^{b'} F^* A \phi_k \, d\xi \right] \left[\int_a^{b'} \phi_k^* A \phi_k \, d\xi \right]^{-1} \left[\int_a^{b} \phi_k^* A F \, d\xi \right].$$

This is Parseval's equality.

DEFINITION 13.3. Let R_k^2 denote the $n \times n$ matrix $[\int_a^{b'} \phi_k^* A \phi_k \, d\xi]^{-1}$.

DEFINITION 13.4. Let

$$G(\lambda) = \int_a^{b'} \phi(\xi, \lambda)^* A(\xi) F(\xi) \, d\xi.$$

DEFINITION 13.5. Let $P_{b'}(\lambda)$ be an $n \times n$ matrix valued function satisfying

(1) $P_{b'}(0+) = 0$;

(2) $P_{b'}(\lambda)$ is increasing, jumping R_k^2 at $\lambda = \lambda_k$, but otherwise constant, continuous from above.

Thus

$$P_{b'}(\lambda) = \sum_{0 < \lambda_k \leqq \lambda} R_k^2, \qquad \lambda \geqq 0,$$

$$P_{b'}(\lambda) = -\sum_{\lambda < \lambda_k \leqq 0} R_k^2, \qquad \lambda < 0.$$

THEOREM 13.6. (Parseval's equality.) *Let F be an arbitrary element of $L_A^2(a, b')$. Then*

$$\int_a^{b'} F^* A F \, d\xi = \int_{-\infty}^{\infty} G^*(\lambda) \, dP_{b'}(\lambda) G(\lambda).$$

This can be extended by the polarization identities to inner products.

COROLLARY 13.7. *Let F_1, F_2 be arbitrary elements in $L_A^2(a, b')$. Let $G_1(\lambda)$ and $G_2(\lambda)$ correspond to them according to Definition* 13.4. *Then*

$$\int_a^{b'} F_2^* A F_1 \, d\xi = \int_{-\infty}^{\infty} G_2^*(\lambda) \, dP_{b'}(\lambda) G_1(\lambda).$$

We will need this in the theorem that follows.

THEOREM 13.8. *There exists a nondecreasing $n \times n$ matrix valued function $P(\lambda)$, defined on $(-\infty, \infty)$, such that*

(1) $P(0+) = 0$,

(2) $P(\lambda) - P(\mu) = \lim_{b' \to b} [P_{b'}(\lambda) - P_{b'}(\mu)]$, $\lambda > \mu$.

Proof. Let $M_{b'}$ be on the circle $E(M)$, defined by setting $\chi_{b'}^*(b', \lambda)(J/i)\chi_b(b', \lambda) = 0$. Set λ equal to λ_0, and $\chi_{b'}(x, \lambda_0) = \theta(x, \lambda_0) + \phi(x, \lambda_0)M_{b'}(\lambda_0)$. Then $J\chi_{b'}' = (\lambda_0 A + B)\chi_{b'}$. Apply Parseval's equality to $\chi_{b'}$.

$$\int_a^{b'} \chi_{b'}^* A \chi_{b'}\, d\xi = \int_{-\infty}^{\infty} G(\lambda)^*\, dP_{b'}(\lambda) G(\lambda)$$

where

$$G(\lambda) = \int_a^{b'} \phi^*(\xi, \lambda) A(\xi) \chi_{b'}(\xi, \lambda_0)\, d\xi.$$

Now $J\chi_b' = (\lambda_0 A + B)\chi_{b'}$ and $J\phi_k' = (\lambda_k A + B)\phi_k$ imply

$$\phi_k^* J \chi_{b'}\Big|_a^{b'} = (\lambda_0 - \lambda_k) \int_a^{b'} \phi_k^* A \chi_{b'}\, d\xi.$$

At b' both ϕ_k and $\chi_{b'}$ satisfy $(\beta_1, \beta_2)\, Y(b') = 0$, so the upper limit on the left is 0. At $x = a$,

$$\phi_k^* J \chi_{b'}(a) = (-\alpha_2, \alpha_1) \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} \alpha_1^* & -\alpha_2^* \\ \alpha_2^* & \alpha_1 \end{pmatrix} \begin{pmatrix} I_n \\ M_{b'} \end{pmatrix} = I_n.$$

So

$$G(\lambda_k) = \int_a^{b'} \phi_k^* A \chi_{b'}\, d\xi = (\lambda_k - \lambda_0)^{-1} I_n.$$

Parseval's equality becomes

$$\int_a^{b'} \chi_{b'}^* A \chi_{b'}\, d\xi = \int_{-\infty}^{\infty} |\lambda - \lambda_0|^{-2}\, dP_{b'}(\lambda).$$

An easy calculation using the differential equation for $\chi_{b'}$ shows

$$\int_a^{b'} \chi_{b'}^* A \chi_{b'}\, d\xi = [M_{b'} - M_{b'}^*]/(2i \operatorname{Im} \lambda_0).$$

If $\lambda_0 = i = \sqrt{-1}$, then

$$\int_{-\infty}^{\infty} |i - \lambda|^{-2}\, dP_{b'}(\lambda) = [M_{b'}(i) - M_{b'}^*(i)]/(2i).$$

Since $|i - \lambda|^2 = \lambda^2 + 1$, we see there is a $K > 0$ such that

$$\int_{-\infty}^{\infty} (\lambda^2 + 1)^{-1}\, dP_{b'}(\lambda) < K.$$

This implies for $\mu > 0$ that

$$\int_{-\mu}^{\mu} dP_{b'}(\lambda) < K[1 + \mu^2].$$

This implies since $P_{b'}$ is increasing, and $0 \leqq P_{b'}(\mu) < K[1 + \mu^2]$, then $0 \leqq -P_{b'}(-\mu) < K[1 + \mu^2]$ for all $\mu \geqq 0$. Therefore $P_{b'}(\lambda)$ is uniformly bounded on compact subintervals of the real line. Helly's first theorem [23] shows there is a subsequence that converges "weakly" to $P(\lambda)$ with the properties stated.

THEOREM 13.9. *If F is in $L_A^2(a, b)$, there is a function $G(\lambda)$ in $L_p^2(-\infty, \infty)$, with inner product*

$$(G, H)_p = \int_{-\infty}^{\infty} H^* \, dPG,$$

such that if

$$E(\lambda) = G(\lambda) - \int_a^{b'} \phi^*(\xi, \lambda) A(\xi) F(\xi) \, d\xi,$$

then

$$\lim_{b' \to b} \int_{-\infty}^{\infty} E(\lambda)^* \, dP(\lambda) E(\lambda) = 0,$$

and

$$\int_a^b F(\xi)^* A(\xi) F(\xi) \, d\xi = \int_{-\infty}^{\infty} G(\lambda)^* \, dP(\lambda) G(\lambda).$$

Proof. (1) Let F be in $C_0^1(a, b)$. Then F is in D.
If b' is large enough,

$$\int_a^{b'} (LF)^* A(LF) = \int_a^{b'} H^* AH \, d\xi = \sum_{k=1}^{\infty} \left(\int_a^b \phi_k^* AH \, d\xi \right)^* R_k^2 \left(\int_a^b \phi_k^* AH \, d\xi \right).$$

But

$$\int_a^b \phi_k^* AH \, d\xi = \int_a^{b'} \phi_k^* [JF' - BF] \, d\xi = \int_a^{b'} [J\phi_k' - B\phi_k]^* F \, d\xi$$

$$= \lambda_k \int_a^{b'} \phi_k^* AF \, d\xi = \lambda_k G(\lambda_k).$$

Thus

$$\int_a^b (LF)^* A(LF) \, d\xi = \int_{-\infty}^{\infty} \lambda^2 G(\lambda)^* \, dP_{b'}(\lambda) G(\lambda).$$

Let $N > 0$. Then

$$\left(\int_{-\infty}^{-N} + \int_N^{\infty} \right) G^* \, dP_{b'} G \leqq \frac{1}{N^2} \left(\int_{-\infty}^{-N} + \int_N^{\infty} \right) \lambda^2 G^* \, dP_{b'} G$$

$$\leqq \frac{1}{N^2} \int_{-\infty}^{\infty} \lambda^2 G^* \, dP_{b'} G \leqq \frac{1}{N^2} \int_a^b (LF)^* A(LF) \, d\xi.$$

So, since

$$\int_a^b F^* AF \, d\xi = \left(\int_{-\infty}^{-N} + \int_{-N}^{N} + \int_N^{\infty} \right) G^* \, dP_{b'} G,$$

$$\left| \int_a^b F^* AF \, d\xi - \int_{-N}^{N} G^* \, dP_{b'} G \right| = \left(\int_{-\infty}^{-N} + \int_N^{\infty} \right) G^* \, dP_{b'} G$$

$$\leqq \frac{1}{N^2} \int_a^b (LF)^* A(LF) \, d\xi.$$

Let b' approach b. Helly's second theorem [23] implies

$$\left| \int_a^b F^* AF\, d\xi - \int_{-N}^{N} G^*\, dPg \right| \leqq \frac{1}{N^2} \int_a^b (LF)^* A(LF)\, d\xi.$$

Let N approach ∞.

$$\int_a^b F^* AF\, d\xi = \int_{-\infty}^{\infty} G^*\, dPG,$$

provided F is in $C_0^1(a, b)$.

(2) Let F vanish near b but otherwise be arbitrary in $L_A^2(a, b)$. Choose $\{F_j\}_{j=1}^{\infty}$ in $C_0^1(a, b)$ such that

$$\lim_{j\to\infty} \int_a^b (F_j - F)^* A(F_j - F)\, d\xi = 0.$$

Apply Parseval's equality to $F_j - F_k$,

$$\int_a^b (F_j - F_k)^* A(F_j - F_k)\, d\xi = \int_{-\infty}^{\infty} (G_j - G_k)^*\, dP(G_j - G_k),$$

where

$$G_j = \int_a^b \phi^* AF_j\, d\xi, \qquad G_k = \int_a^b \phi^* AF_k\, d\xi.$$

Since $\lim_{j\to\infty} F_j = F$, $\{G_j\}_{j=1}^{\infty}$ is also a Cauchy sequence in $L_p^2(-\infty, \infty)$. Thus there is a G in $L_p^2(-\infty, \infty)$ such that $\lim_{j\to\infty} G_j = G$. Since

$$\left| G_j - \int_a^{b'} \phi^*\, AF\, d\xi \right| = \left| \int_a^{b'} \phi^* A(F_j - F)\, d\xi \right|$$

$$\leqq \left(\int_a^{b'} \phi^* A\phi\, d\xi \right)^{1/2} \left(\int_a^{b'} [F_j - F]^* A[F_j - F]\, d\xi \right)^{1/2}$$

implies $\lim_{j\to\infty} G_j = \int_a^{b'} \phi^* AF\, d\xi$, which is continuous, we find $G = \int_a^b \phi^* AF\, d\xi$, almost everywhere. Thus if F vanishes near $x = b$,

$$\int_a^b F^* AF\, d\xi = \lim_{j\to\infty} \int_a^b F_j^* AF_j\, d\xi = \lim_{j\to\infty} \int_{-\infty}^{\infty} G_j^*\, dPG_j = \int_{-\infty}^{\infty} G^*\, dPG.$$

(3) Finally, if F is arbitrary in $L_A^2(a, b)$, let

$$F_{b'} = F, \qquad x \leqq b',$$

$$F_{b'} = 0, \qquad x > b'.$$

Let

$$G_{b'} = \int_a^b \phi^* AF_{b'}\, d\xi = \int_a^{b'} \phi^* AF\, d\xi.$$

Since

$$\int_{-\infty}^{\infty} (G_c - G_d)^*\, dP(G_c - G_d) = \int_c^d F^* AF\, d\xi,$$

$\{G_{b'}\}$ is a Cauchy sequence as b' approaches b. Let $\lim_{b'\to b} G_{b'} = G$ in $L_p^2(-\infty, \infty)$. Letting b' approach b in the previous result, we get

$$\int_a^b F^*AF\,d\xi = \int_{-\infty}^{\infty} G^*\,dPG.$$

(4) Since $G_{b'}$ approaches G in $L_p^2(-\infty, \infty)$,

$$\lim_{b'\to b} \int_{-\infty}^{\infty} \left[G(\lambda) - \int_a^{b'} \phi^*AF\,d\xi\right]^* dP(\lambda) \left[G(\lambda) - \int_a^{b'} \phi^*AF\,d\xi\right] = 0.$$

We remark that if the condition $JY' - BY = AF$, $AY = 0$ implies $Y = 0$ fails to hold, then the approximation used in the proof of this theorem may hold only on a subspace of $L_A^2(a, b)$. The subspace may be as small as only one dimension. We will give examples later.

THEOREM 13.10. *If $G(\lambda)$ is the limit of $\int_a^{b'} \phi(\xi, \lambda)^*A(\xi)F(\xi)\,d\xi$ in $L_p^2(-\infty, \infty)$, then*

$$\int_{-\infty}^{\infty} \phi(x, \lambda)\,dP(\lambda)G(\lambda) = F(x)$$

in $L_A^2(a, b)$; that is,

$$\lim_{I\to(-\infty,\infty)} \int_a^b \left[F - \int_I \phi\,dPG\right]^* A \left[F - \int_I \phi\,dPG\right] d\xi = 0.$$

Proof. Let $I = (\mu, \nu)$, and

$$F_I(x) = \int_I \phi(x, \lambda)\,dP(\lambda)G(\lambda).$$

If b' is in $[a, b)$, then

$$\int_a^{b'} [F - F_I]^*AF_I\,d\xi = \int_a^{b'} [F - F_I]^*A\left[\int_I \phi\,dPG\right] d\xi$$
$$= \int_I \left[\int_a^{b'} [F - F_I]^*A\phi\,d\xi\right] dPG.$$

Likewise, we have

$$\int_a^{b'} [F - F_I]^*AF\,d\xi = \int_{-\infty}^{\infty} \left[\int_a^{b'} [F - F_I]^*A\phi\,d\xi\right] dPG.$$

Subtracting, we obtain

$$\int_a^{b'} [F - F_I]^*A[F - F_I]\,d\xi = \int_{(-\infty,\infty)-I} \left[\int_a^{b'} [F - F_I]^*A\phi\,d\xi\right] dPG.$$

Now $\int_a^{b'} \phi^*A[F - F_I]\,d\xi$ is the transform of a function in $L_A^2(a, b)$, which vanishes on (b', b). Consequently the integral from a to b' in brackets is in $L_p^2(-\infty, \infty)$. Applying Schwarz's inequality,

$$\left(\int_a^{b'} [F - F_I]^*A[F - F_I]\,d\xi\right)^2$$
$$\leqq \left(\int_{(-\infty,\infty)-I} \left[\int_a^{b'} \phi^*A[F - F_I]\,d\xi\right]^* dP \left[\int_a^{b'} \phi^*A[F - F_I]\,d\xi\right]\right)$$
$$\times \left(\int_{(-\infty,\infty)-I} G^*\,dPG\right).$$

The first integral on the right is less than or equal to

$$\int_a^{b'} [F-F_I]^* A[F-F_I]\, d\xi.$$

If this is inserted and canceled,

$$\int_a^{b'} [F-F_I]A[F-F_I]\, d\xi \leqq \int_{(-\infty,\infty)-I} G^*\, dPG.$$

Let b' approach b. Then let I approach $(-\infty, \infty)$. The result is that $F = \lim_{I\to(-\infty,\infty)} F_I$, or

$$F(x) = \lim_{(\mu,\nu)\to(-\infty,\infty)} \int_\mu^\nu \phi(x,\lambda)\, dP(\lambda)G(\lambda)$$

in $L_A^2(a, b)$.

Theorem 13.9 may be extended to involve inner products by use of the polarization identity. The inner product form of Parseval's equality is

$$\int_a^b F_2(\xi)^* A(\xi)F_1(\xi)\, d\xi = \int_{-\infty}^{\infty} G_2(\lambda)^*\, dP(\lambda)G_1(\lambda),$$

where

$$G_j(\lambda) = \int_a^b \phi(\xi,\lambda)^* A(\xi)F_j(\xi)\, d\xi, \qquad j = 1, 2.$$

Theorems 13.9 and 13.10 may be extended to represent the resolvent operator $(L-\lambda_0 I)^{-1}$ when λ_0 is not in the support of $dP(\lambda)$. Parseval's equality is

$$\int_a^b [(L-\lambda_0 I)^{-1}F(\xi)]^* A(\xi)[(L-\lambda_0 I)^{-1}F(\xi)]\, d\xi = \int_{-\infty}^{\infty} \frac{G(\lambda)^*\, dP(\lambda)G(\lambda)}{|\lambda-\lambda_0|^2}.$$

The resolvent expansion is

$$(L-\lambda_0 I)^{-1}F(x) = \int_{-\infty}^{\infty} \phi(x,\lambda)\, dP(\lambda)\frac{G(\lambda)}{\lambda-\lambda_0}.$$

14. The converse problem. Again, this section follows closely the lead of Coddington and Levinson [3, Chap. 9].

The preceding section began with choosing an F in $L_A^2(a, b)$, producing G in $L_p^2(-\infty, \infty)$, and then showing that F could be recovered from G. In this section we begin with G, produce F, and then recover G.

Without the assumption that $JY' - BY = 0$, $AY = 0$ implies $Y = 0$, $L_p^2(-\infty, \infty)$ may be too large in the sense that $G \to F \to \tilde{G}$, but $\tilde{G}$ may not equal G. $\tilde{G}$ may be only in a subspace of $L_p^2(-\infty, \infty)$.

With the assumption made in the Introduction (let $F = 0$), there is no difficulty.

LEMMA 14.1. *Let $G(\lambda)$ be in $L_p^2(-\infty, \infty)$. Let*

$$F_I(x) = \int_I \phi(x,\lambda)\, dP(\lambda)G(\lambda).$$

Then $\lim_{I\to(-\infty,\infty)} F_I(x)$ exists in $L_A^2(a, b)$.

Proof. Let $I_1 \subset I_2$. Then

$$F_{I_2} - F_{I_1} = \int_{I_2-I_1} \phi(x,\lambda)\, dP(\lambda)G(\lambda) = \int_{-\infty}^{\infty} \phi(x,\lambda)\, dP(\lambda)K_{I_2-I_1}(\lambda)G(\lambda)$$

where

$$K_{I_2-I_1}(\lambda)=1, \qquad \lambda\in I_2-I_1,$$
$$K_{I_2-I_1}(\lambda)=0, \qquad \lambda\notin I_2-I_1.$$

Let F be an arbitrary element of $L_A^2(a,b)$, which vanishes near b. Then

$$\int_a^b R^*A[F_{I_2}-F_{I_1}]\,d\xi=\int_a^b R^*A\left[\int_{I_2-I_1}\phi\,dPG\right]d\xi$$
$$=\int_{I_2-I_1}\left[\int_a^b \phi^*AR\,d\xi\right]^*dPG=\int_{I_2-I_1}S^*\,dPG,$$

where S is the transform of R: $S=\int_a^b \phi^*AR\,d\xi$. We now let $R=F_{I_2}-F_{I_1}$ on $[a,b')$, but set $R=0$ near b. Then

$$\int_a^b [F_{I_2}-F_{I_1}]^*A[F_{I_2}-F_{I_1}]\,d\xi=\int_{I_2-I_1}S^*\,dPG$$
$$\leqq\left(\int_{I_2-I_1}S^*\,dPS\right)^{1/2}\left(\int_{I_2-I_1}G^*\,dPG\right)^{1/2}$$
$$\leqq\left(\int_{-\infty}^{\infty}S^*\,dPS\right)^{1/2}\left(\int_{I_2-I_1}G^*\,dPG\right)^{1/2}.$$

But by Parseval's equality

$$\int_a^b [F_{I_2}-F_{I_1}]^*A[F_{I_2}-F_{I_1}]\,d\xi=\int_{-\infty}^{\infty}S^*\,dPS.$$

Take the square root, cancel with the inequality above, and then square.

$$\int_a^b [F_{I_2}-F_{I_1}]^*A[F_{I_2}-F_{I_1}]\,d\xi\leqq\int_{I_2-I_1}G^*\,dPG.$$

Since the right-hand side is independent of b', let b' approach b. Then $F_{I_2}-F_{I_1}$ is in $L_A^2(a,b)$. As I approaches $(-\infty,\infty)$, the inequality also shows that $\{F_I\}$ forms a Cauchy sequence in $L_A^2(a,b)$, and therefore $\lim_{I\to(-\infty,\infty)}F_I=F$ in $L_A^2(a,b)$.

LEMMA 14.2. *Let $G(\lambda)$ be in $L_P^2(-\infty,\infty)$. Let*

$$F_I(x)=\int_I\phi(x,\lambda)\,dP(\lambda)G(\lambda),$$

$$F(x)=\lim_{I\to(-\infty,\infty)}\int_I\phi(x,\lambda)\,dP(\lambda)G(\lambda).$$

Let

$$\tilde{G}(\lambda)=\int_a^b\phi^*(\xi,\lambda)A(\xi)F(\xi)\,d\xi,$$

and let

$$\tilde{F}_I(x)=\int_I\phi(x,\lambda)\,dP(\lambda)\tilde{G}(\lambda).$$

Then

$$\lim_{I\to(-\infty,\infty)}\int_a^b[F_I-\tilde{F}_I]^*A[F_I-\tilde{F}_I]\,d\xi=0.$$

Proof. From Theorem 13.10, $F(x) = \lim_{I\to(-\infty,\infty)} \tilde{F}_I(x)$ in $L^2_A(a, b)$. But by definition, $F(x) = \lim_{I\to(-\infty,\infty)} F_I(x)$. The triangle inequality

$$\|F_I - \tilde{F}_I\|_A \leqq \|F_I - F\|_A + \|F - \tilde{F}_I\|_A$$

shows that as I approaches $(-\infty, \infty)$, $\|F_I - \tilde{F}_I\|_A$ approaches zero.

At this point we have, given G, that an F exists. F yields $\tilde{G}$ which again yields F. And so the process stops. We continue to show G and $\tilde{G}$ coincide.

LEMMA 14.3. *Let λ_0 be a complex number with positive imaginary part, and let*

$$H_I(x, \lambda_0) = \int_I \phi(x, \lambda)\, dP(\lambda) \left[\frac{G(\lambda) - \tilde{G}(\lambda)}{\lambda - \lambda_0}\right].$$

Then for all fixed λ_0, $\lim_{I\to(-\infty,\infty)} H_I(x, \lambda_0) = 0$.

Proof. H_I satisfies

$$JH_I' - [\lambda_0 A + B]H_I = A(x) \int_I [\phi(x, \lambda)\, dP(\lambda)][G(\lambda) - \tilde{G}(\lambda)].$$

Further, H_I satisfies the boundary condition $(\alpha_1, \alpha_2)H_I(a, \lambda_0) = 0$. Thus

$$H_I(x) = \int_a^b G(\lambda_0, x, \xi)A(\xi)[F_I(\xi) - \tilde{F}_I(\xi)]\, d\xi + \phi(x, \lambda_0)C,$$

where here $G(\lambda_0, x, \xi)$ is the Green's function for the singular boundary value problem. As I approaches $(-\infty, \infty)$, the integral approaches zero. So

$$H(x, \lambda_0) = \lim_{I\to(-\infty,\infty)} H_I(x, \lambda_0) = \phi(x, \lambda_0)C.$$

But $\lim_{x\to b} \chi(x, \bar{\lambda}_0)^* JH(x, \lambda_0) = 0$ because, since H_I satisfies the boundary condition as x approaches b, H must satisfy the singular boundary condition as x approaches b as well. Since $\lim_{x\to b} \chi(x, \bar{\lambda}_0)^* J\phi(x, \lambda_0) = M(\lambda)$, which is nonsingular, $C = 0$.

LEMMA 14.4. *$G(\lambda) = \tilde{G}(\lambda)$ in $L^2_p(-\infty, \infty)$.*

Proof. Let K be a constant $2n \times 1$ matrix; let

$$Y_s(\lambda) = \int_a^s \phi(\xi, \lambda)^* A(\xi) K\, d\xi.$$

Then $Y_s(\lambda)$ is in $L^2_p(-\infty, \infty)$, since it is the transform of a function that vanishes near b. Then

$$\int_a^s K^* A(\xi) H_I(\xi, \lambda_0)\, d\xi = \int_I Y_s(\lambda)^*\, dP(\lambda) \left[\frac{G(\lambda) - \tilde{G}(\lambda)}{\lambda - \lambda_0}\right].$$

Let I approach $(-\infty, \infty)$. Then

$$0 = \int_{-\infty}^{\infty} Y_s(\lambda)^*\, dP(\lambda) \left[\frac{G(\lambda) - \tilde{G}(\lambda)}{\lambda - \lambda_0}\right].$$

By varying μ_0 and ν_0 ($\lambda_0 = \mu_0 + i\nu_0$) independently, we see that

$$0 = \int_{-\infty}^{\infty} Y_s(\lambda)^*\, dP(\lambda)[G(\lambda) - \tilde{G}(\lambda)] \left[\frac{\nu_0}{(\lambda - \mu_0)^2 + \nu_0^2}\right].$$

If we integrate with respect to μ_0 from α to β, reversing the order of integration, we have

$$0 = \int_{-\infty}^{\infty} Y_s(\lambda)^*\, dP(\lambda)[G(\lambda) - \tilde{G}(\lambda)] \left[\tan^{-1}\left(\frac{\beta - \lambda}{\nu_0}\right) - \tan^{-1}\left(\frac{\alpha - \lambda}{\nu_0}\right)\right].$$

Letting ν_0 approach zero, we get

$$0=\int_\alpha^\beta Y_s(\lambda)^*\, dP(\lambda)[G(\lambda)-\tilde{G}(\lambda)],$$

or

$$0=\int_a^s\left[\int_\alpha^\beta K^*A(\xi)\phi(\xi,\lambda)\, dP(\lambda)[G(\lambda)-\tilde{G}(\lambda)]\right]d\xi.$$

Differentiate with respect to s:

$$0=\int_\alpha^\beta K^*A(x)\phi(x,\lambda)\, dP(\lambda)[G(\lambda)-\tilde{G}(\lambda)].$$

Now apply an extension of the Mean Value Theorem, remembering that the expression is 1×1 and ϕ is analytic in λ. We find for some λ_0 in $[\alpha,\beta]$

$$K^*A(x)\phi(x,\lambda_0)\int_\alpha^\beta dP(\lambda)[G(\lambda)-\tilde{G}(\lambda)]=0.$$

Since $A(x)\phi(x,\lambda_0)v$ is only zero for all x when $v=0$, we can choose K appropriately to conclude that

$$\int_\alpha^\beta dP(\lambda)[G(\lambda)-\tilde{G}(\lambda)]=0$$

for all α,β. We use this to build up integrals involving step functions, dense in $L_p^2(-\infty,\infty)$, which have as their limit

$$\int_{-\infty}^{\infty}[G(\lambda)-\tilde{G}(\lambda)]^*\, dP(\lambda)[G(\lambda)-\tilde{G}(\lambda)]=0.$$

Hence $G=\tilde{G}$ in $L_p^2(-\infty,\infty)$.

We summarize in the following theorem.

THEOREM 14.5. *If $G(\lambda)$ is in $L_p^2(-\infty,\infty)$, there is a unique $F(x)$ in $L_p^2(a,b)$ such that*

$$F(x)=\int_{-\infty}^{\infty}\phi(x,\lambda)\, dP(\lambda)G(\lambda)$$

and

$$G(\lambda)=\int_a^b\phi(\xi,\lambda)^*A(\xi)F(\xi)\, d\xi.$$

15. The relation between $M(\lambda)$ and $P(\lambda)$. The matrix $M(\lambda)$ can frequently be determined by careful inspection of the solutions of (*) to determine appropriate solutions in $L_A^2(a,b)$. More difficult is the determination of the spectral measure $P(\lambda)$, since its existence follows from Helly's selection theorems. Fortunately, they are intimately connected. Recall the following theorems.

THEOREM 15.1. *Let $M(\lambda)$ be on the limit circle; let $\chi(x,\lambda)=\theta(x,\lambda)+\phi(x,\lambda)M(\lambda)$ for $\lambda=\mu+i\nu$, $\nu\neq 0$. Then*

$$\int_a^b\chi(\xi,\lambda)^*A(\xi)\chi(\xi,\lambda)\, d\xi=[M(\lambda)-M(\lambda)^*]/(2i\ \mathrm{Im}\,\lambda).$$

THEOREM 15.2. *Let* $\lambda_0 = \mu + i\nu$, $\nu \neq 0$. *Then*

$$\int_a^b \chi(\xi, \lambda_0)^* A(\xi)\chi(\xi, \lambda_0)\, d\xi = \int_{-\infty}^{\infty} |\lambda - \lambda_0|^{-2}\, dP(\lambda).$$

Proof. The theorem is true if b is replaced by b'. Let b' approach b.

THEOREM 15.3. *If* λ_1 *and* λ_2 *are real, then*

$$P(\lambda_2) - P(\lambda_1) = \lim_{\nu \to 0+} \frac{1}{\pi} \int_{\lambda_1}^{\lambda_2} \operatorname{Im} M(\mu + i\nu)\, d\mu.$$

Further, if λ_1 *and* λ_2 *have nonzero imaginary parts, then*

$$M(\lambda_2) - M(\lambda_1) = \int_{-\infty}^{\infty} [(\lambda - \lambda_2)^{-1} - (\lambda - \lambda_1)^{-1}]\, dP(\lambda).$$

Proof. We have $\int_{-\infty}^{\infty} |\lambda - \lambda_0|^{-2}\, dP(\lambda) = [M(\lambda_0) - M(\lambda_0)^*]/2i$ satisfies

$$\operatorname{Im} M(\mu + i\nu) = \int_{-\infty}^{\infty} \frac{\nu\, dP(\lambda)}{(\lambda - \mu)^2 + \nu^2}.$$

Integrate both sides from λ_1 to λ_2 with respect to μ.

$$\int_{\lambda_1}^{\lambda_2} \operatorname{Im} M(\mu + i\nu)\, d\mu = \int_{\lambda_1}^{\lambda_2} \int_{-\infty}^{\infty} \frac{\nu\, dP(\lambda)}{(\lambda - \mu)^2 + \nu^2}\, d\mu$$

$$= \int_{-\infty}^{\infty} \left[\tan^{-1}\left(\frac{\lambda_2 - \lambda}{\nu}\right) - \tan^{-1}\left(\frac{\lambda_1 - \lambda}{\nu}\right) \right] dP(\lambda).$$

Let ν approach 0 from above, we obtain

$$\lim_{\nu \to 0+} \int_{\lambda_1}^{\lambda_2} \operatorname{Im} M(\mu + i\nu)\, d\mu = \pi \int_{\lambda_1}^{\lambda_2} dP(\lambda) = \pi[P(\lambda_2) - P(\lambda_1)].$$

To validate the second part, note

$$\operatorname{Im}[M(\lambda_0) - M(\lambda_1)] = \int_{-\infty}^{\infty} \left[\frac{\nu_2}{|\lambda - \lambda_2|^2} - \frac{\nu_1}{|\lambda - \lambda_1|^2} \right] dP(\lambda)$$

$$= \operatorname{Im} \int_{-\infty}^{\infty} \left[\frac{(\lambda - \lambda_2)^*}{|\lambda - \lambda_2|^2} - \frac{(\lambda - \lambda_1)^*}{|\lambda - \lambda_1|^2} \right] dP(\lambda)$$

$$= \operatorname{Im} \int_{-\infty}^{\infty} [(\lambda - \lambda_2)^{-1} - (\lambda - \lambda_1)^{-1}]\, dP(\lambda).$$

We therefore have two functions, analytic in λ_1 and λ_2, whose imaginary parts are equal. The real parts can only differ by a constant. Letting $\lambda_2 = \lambda_1$ shows this constant to be zero.

16. The spectral resolution. We connect the result of the preceding sections to the representation of the identity as an integral generated by a projection valued measure E_λ. Given

$$F(x) = \int_{-\infty}^{\infty} \phi(x, \lambda)\, dP(\lambda) G(\lambda)$$

where

$$G(\lambda) = \int_a^b \phi^*(\xi, \lambda) A(\xi) F(\xi)\, d\xi,$$

we define

$$E_\lambda F(x) = \int_{-\infty}^{\lambda_+} \phi(x, \lambda)\, dP(\lambda) G(\lambda).$$

Considered the limit of eigenfunction expansions, E_λ can easily be shown to be a projection. It is continuous from above and satisfies $E_{\lambda_1} E_{\lambda_2} = E_{\lambda_1}$ when $\lambda_1 \leqq \lambda_2$, as well as $E_{-\infty} = O$, $E_\infty = I$. If we let $\{\lambda_j\}_{j=-\infty}^{\infty}$ be a partition of $(-\infty, \infty)$, $\lambda_i < \lambda_j$ if $i < j$, and

$$\Delta_j E F(x) = \int_{\lambda_j}^{\lambda_{j+1}} \phi(x, \lambda)\, dP(\lambda) G(\lambda),$$

then $F(x) = \sum_{j=-\infty}^{\infty} \Delta_j E F(x)$. As the mesh $\{\lambda_j\}_{j=-\infty}^{\infty}$ becomes finer, we may write

$$F(x) = \int_{-\infty}^{\infty} dE_\lambda F(x)$$

as the limit of the decomposition above.

If Y is in D, it has the representation

$$Y(x) = \int_{-\infty}^{\infty} \phi(x, \lambda)\, dP(\lambda) G(\lambda)$$

where

$$G(\lambda) = \int_a^b \phi^*(\xi, \lambda) A(\xi) Y(\xi)\, d\xi.$$

Then

$$LY(x) = \int_{-\infty}^{\infty} \lambda \phi(x, \lambda)\, dP(\lambda) G(\lambda).$$

This translates into

$$LY(x) = \int_{-\infty}^{\infty} \lambda\, dE_\lambda Y(x).$$

The resolvent operator also has the standard representation. If λ_0 is complex,

$$(L - \lambda_0 I)^{-1} F(x) = \int_{-\infty}^{\infty} \frac{1}{\lambda - \lambda_0}\, dE_\lambda F(x).$$

It is apparent that λ_0 is in the spectrum of L if and only if it is in the support of dE_λ or $dP(\lambda)$.

17. Examples. (1) Consider the two-dimensional system

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \left[\lambda \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & (x^2 - 1)^{-1} \end{pmatrix} \right] \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

on the interval $[0, 1)$, together with boundary conditions

$$O: \quad (1, 0) \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = 0$$

or

$$E: \quad (0, 1) \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = 0$$

at $x=0$, and

$$\lim_{x\to 1}(1,0)\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}y_1(x)\\ y_2(x)\end{pmatrix}=0$$

at $x=1$.

The differential system is equivalent to the Legendre differential equation. With the O boundary condition, the odd polynomial boundary value problem is generated. With the E boundary condition, the even polynomial boundary value problem arises. The boundary condition at $x=1$ is the one satisfied by the Legendre polynomials. $x=1$ is limit circle.

(2) Consider the four-dimensional system

$$\begin{pmatrix}0&0&-1&0\\0&0&0&-1\\1&0&0&0\\0&1&0&0\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}'=\left[\lambda\begin{pmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix}+\begin{pmatrix}0&0&0&0\\0&-2(1-x^2)&1&0\\0&0&0&0\\0&0&0&(1-x^2)^{-2}\end{pmatrix}\right]\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}$$

on the interval $[0,1)$, together with boundary conditions

$$O:\quad (1,0,0,0)\begin{pmatrix}y_1(0)\\y_2(0)\\y_3(0)\\y_4(0)\end{pmatrix}=0,\qquad (0,0,0,1)\begin{pmatrix}y_1(0)\\y_2(0)\\y_3(0)\\y_4(0)\end{pmatrix}=0$$

or

$$E:\quad (0,1,0,0)\begin{pmatrix}y_1(0)\\y_2(0)\\y_3(0)\\y_4(0)\end{pmatrix}=0,\qquad (0,0,1,0)\begin{pmatrix}y_1(0)\\y_2(0)\\y_3(0)\\y_4(0)\end{pmatrix}=0$$

at $x=0$, and

$$\lim_{x\to 1}(\ln(1+x),(1+x)^{-1},0,-(1-x)^2)\begin{pmatrix}0&0&-1&0\\0&0&0&-1\\1&0&0&0\\0&1&0&0\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=0,$$

$$\lim_{x\to 1}(1,0,0,0)\begin{pmatrix}0&0&-1&0\\0&0&0&-1\\1&0&0&0\\0&1&0&0\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=0$$

where, at $x=1$, the limit-4 case holds, and for $\lambda=0$, $(1,0,0,0)^T$, and $(\ln(1+x),(1+x)^{-1},0,-(1-x)^2)^T$ are L^2 solutions. The problem with O boundary conditions is again the odd-degree Legendre polynomial boundary value problem. With E boundary conditions, the even-degree Legendre polynomial value problem is the result.

(3) Consider the fourth-order scalar problem $(y'')''=\lambda y$ on $[0,\infty)$, together with boundary conditions $y(0)=0$, $y'''(0)=0$. ∞ is limit point, so no boundary condition at ∞ is required.

Let $y_1 = y$, $y_2 = y'$, $y_3 = -y'''$, $y_4 = y''$; then the problem is equivalent to

$$\begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}' = \left[\lambda \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right] \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix},$$

with boundary condition

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} (0) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The initial condition for the fundamental matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The components of the four 4×1 matrices that makes its columns are combinations of $\sinh z_0 x$, $\cosh z_0 x$, and $\cos z_0 x$, where $z_0 = \lambda^{1/4}$. Multiplying the fundamental matrix by

$$\begin{pmatrix} I_2 \\ M \end{pmatrix}$$

and requiring the result to be in $L_A^2(0, \infty)$ shows that

$$M = \begin{pmatrix} z_0^3(1+i)/2 & z_0(1-i)/2 \\ z_0(-i+1)/2 & (1+i)/2z_0 \end{pmatrix}.$$

As $\lambda = z_0^4$ approaches the positive real axis,

$$\operatorname{Im} M \to \begin{pmatrix} \lambda^{3/4}/2 & -\lambda^{1/4}/2 \\ -\lambda^{1/4}/2 & 1/2\lambda^{1/4} \end{pmatrix}.$$

As λ approaches the negative real axis, $\operatorname{Im} M \to 0$. Consequently,

$$dP(\lambda) = \frac{1}{\pi} \begin{pmatrix} \lambda^{3/4}/2 & \lambda^{1/4}/2 \\ -\lambda^{1/4}/2 & 1/2\lambda^{1/4} \end{pmatrix} d\lambda, \qquad \lambda \geqq 0.$$

Elements in $L_A^2(0, \infty)$ are dependent only on their first component, so $F = (f(x), 0, 0, 0)^T$. Then

$$G(\lambda) = \frac{1}{2\lambda^{3/4}} \int_0^\infty \begin{pmatrix} \sin \lambda^{1/4} \xi \\ -\lambda^{1/2} \sin \lambda^{1/4} \xi \end{pmatrix} f(\xi)\, d\xi,$$

$$F(x) = \begin{pmatrix} \dfrac{2}{\pi} \displaystyle\int_0^\infty \sin \mu x \left[\int_0^\infty \sin \mu \xi f(\xi)\, d\xi \right] d\mu \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

where we have replaced λ by μ^4. This is the Fourier sine expansion.

18. Subspace expansions. At the beginning the assumption $JY' - BY = AF$, $AY = 0$ implies $Y = 0$ was made. We would like to briefly comment on what occurs if the assumption fails to hold.

First A must be singular. Thus if F is in the maximal domain: $JF' - BF = AH$, the dimension of AH is less than $2n$. If function $\gamma_1, \cdots, \gamma_{2n}$ is chosen so that

$$(\gamma_1, \cdots, \gamma_{2n})A = 0,$$

then

$$(\gamma_1, \cdots, \gamma_{2n})(JF' - BF) = 0.$$

Letting $(\gamma_1, \cdots, \gamma_{2n})J = K$, $(\gamma_1, \cdots, \gamma_{2n})B = L$, we have $K_jF_j' - L_jF_j = 0, j = 1, \cdots, 2n$.

Three possibilities occur:

(1) If $K_j \neq 0$, $F_j = c_j \exp \int_a^x (L_j/K_j)\, d\xi$.

(2) If $K_j = 0$, $L_j = 0$, F_j is arbitrary.

(3) If $K_j = 0$, $L_j \neq 0$, $F_j = 0$.

Since not all $K_j = 0$, F_j is restricted and may not be dense in $L_A^2(a, b)$.

We provide two examples to illustrate. First consider

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \lambda \begin{pmatrix} 4 & -1 \\ -1 & 1/4 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 4 & -1 \\ -1 & 1/4 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

The fundamental matrix is

$$Y(x, \lambda) = \begin{pmatrix} 1 - \lambda x & 1/4\lambda x \\ -4\lambda x & 1 + \lambda x \end{pmatrix}.$$

Elements in $L_A^2(0, 1)$ have an inner product

$$\left\langle \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \right\rangle = \int_0^1 (\bar{g}_1, \bar{g}_2)\begin{pmatrix} 4 & -1 \\ -1 & 1/4 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} d\xi;$$

an element is zero if $2f_1 = \frac{1}{2}f_2$.

If boundary conditions

$$(1, 0)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}(0) = 0, \qquad (0, 1)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}(1) = 0$$

are imposed, there is only one eigenvalue $\lambda = -1$ and one eigenfunction

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -(1/4)x \\ 1 - x \end{pmatrix}.$$

The solution ϕ is

$$\begin{pmatrix} (1/4)\lambda x \\ 1 + \lambda x \end{pmatrix},$$

so for $F = \binom{f_1}{f_2}$,

$$G(\lambda) = \int_0^1 \left(-f_1 + \frac{1}{4} f_2 \right) d\xi.$$

dP is zero unless $\lambda = -1$. $dP(-1) = 4$. So

$$F(x) = \begin{pmatrix} -(1/4)x \\ 1 - x \end{pmatrix} \cdot 4 \cdot \int_0^1 \left(-f_1 + \frac{1}{4} f_2 \right) d\xi.$$

Clearly this is one-dimensional. In fact, elements in the maximal domain are of the form $\binom{c}{0}$, where c is constant. Such elements are equivalent to the eigenfunction

$$\begin{pmatrix} -(1/4)x \\ 1-x \end{pmatrix}.$$

Only these can be expanded.

Second, consider

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \left[\lambda\begin{pmatrix} e^{-x} & -1 \\ -1 & e^{x} \end{pmatrix}\right]\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} e^{-x} & -1 \\ -1 & e^{x} \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix},$$

$$(1,0)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}(0) = 0, \qquad (0,1)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}(1) = 0.$$

The only eigenvalue is at $\lambda = (1-e)^{-1}$. Its eigenvector is

$$\begin{pmatrix} [e^{x}-1]/[1-e] \\ [1-ee^{-x}]/[1-e] \end{pmatrix}.$$

If

$$\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

is an element of $L_A^2(0,1)$, then

$$G(\lambda) = \int_0^1 (-e^{-\xi}f_1 + f_2)\, d\xi.$$

dP is zero except at $\lambda = (1-e)^{-1}$, where it is $e/(1-e)$. Thus if

$$F(x) = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

is in the maximal domain, D_M,

$$F(x) = \begin{pmatrix} [(e^{x}-1]/[1-e] \\ [1-ee^{x}]/[1-e] \end{pmatrix}\begin{pmatrix} e \\ 1-e \end{pmatrix}\int_0^1 (-e^{-\xi}f_1 + f_2)\, d\xi.$$

A quick calculation shows that elements in D_M have the form $\binom{c}{0}$, where c is constant. Arbitrary elements in $L_A^2(0,1)$ have the form

$$\begin{pmatrix} f_1 - e^{x}f_2 \\ 0 \end{pmatrix},$$

so D_M is not dense.

REFERENCES

[1] F. V. ATKINSON, *Discrete and Continuous Boundary Value Problems*, Academic Press, New York, 1964.

[2] F. BRAUER, *Spectral theory for linear systems of differential equations*, Pacific J. Math., 10 (1960), pp. 17-34.

[3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, II, Wiley-Interscience, New York, 1963.

[5] W. N. EVERITT, *Integrable-square solutions of ordinary differential equations*, Quart. J. Math. Oxford Ser. 2, 10 (1959), pp. 145-155.

[6] W. N. EVERITT, *Integrable-square solutions of ordinary differential equations* II, Quart. J. Math. Oxford Ser. 1, 13 (1962), pp. 217-220.
[7] ———, *Integrable-square solutions of ordinary differential equations* III, Quart. J. Math. Oxford Ser. 1, 14 (1963), pp. 170-180.
[8] ———, *A note on the self-adjoint domains of second-order differential equations*, Quart. J. Math. Oxford Ser. 2, 14 (1963), pp. 41-45.
[9] ———, *Fourth order singular differential equations*, Math. Ann., 149 (1963), pp. 320-340.
[10a] ———, *Singular differential equations* I: *The even order case*, Math. Ann., 156 (1964), pp. 9-24.
[10b] ———, *Singular differential equations* II: *Some self-adjoint even order cases*, Quart. J. Math. Oxford, 18 (1967), pp. 13-32.
[11] ———, *Legendre polynomials and singular differential operators*, in Lecture Notes in Math. 827, Springer-Verlag, New York, Berlin, 1980, pp. 83-106.
[12] W. N. EVERITT AND V. K. KUMAR, *On the Titchmarsh-Weyl theory of ordinary symmetric differential expressions*, I *and* II, Arch. v. Wisk., 3 (1976), pp. 1-48, pp. 109-145.
[13] D. B. HINTON AND J. K. SHAW, *Titchmarsh-Weyl theory for Hamiltonian systems*, in Spectral Theories of Differential Operations, North-Holland, Amsterdam, I. W. Knowles and R. E. Lewis, eds., 1981, pp. 219-230.
[14] ———, *On Titchmarsh-Weyl $M(\lambda)$-functions for linear Hamiltonian systems*, J. Differential Equations, 40 (1981), pp. 316-342.
[15] ———, *On the spectrum of a singular Hamiltonian system*, Quaes. Math., 5 (1982), pp. 29-81.
[16] ———, *Titchmarsh's λ-dependent boundary conditions for Hamiltonian systems*, in Lecture Notes in Math. 964, Springer-Verlag, Berlin, New York, 1982, pp. 318-326.
[17] ———, *Well-posed boundary value problems for Hamiltonian systems of limit point or limit circle type*, in Lecture Notes in Math. 964, Springer-Verlag, Berlin, New York, 1982, pp. 614-631.
[18] ———, *Parameterization of the $M(\lambda)$ function for a Hamiltonian system of limit circle type*, Proc. Roy. Soc. Edinburgh, 93, 1983, pp. 349-360.
[19] ———, *Hamiltonian systems of limit point or limit circle type with both end points singular*, J. Differential Equations, 50 (1983), pp. 444-464.
[20] ———, *On boundary value problems for Hamiltonian systems with two singular points*, SIAM J. Math. Anal., 15 (1984), pp. 272-286.
[21] K. KODAIRA, *On ordinary differential equations of any even order and the corresponding eigenfunction expansions*, Amer. J. Math., 72 (1950), pp. 502-544.
[22] I. KOGAN AND F. S. ROFE-BEKETOV, *On square integrable solutions of symmetric systems of differential equations of arbitrary order*, in Proc. Roy. Soc. Edinburgh, 74, 1974, pp. 5-40.
[23] A. M. KRALL, *Applied Analysis*, D. Reidel, Dordrecht, the Netherlands, 1986.
[24] A. M. KRALL, D. B. HINTON, AND J. K. SHAW, *Boundary conditions for differential systems in intermediate limit situations*, in Proc. Conference Ordinary Partial Differential Equations, I. W. Knowles and R. E. Lewis, eds., North-Holland, Amsterdam, 1984, pp. 301-305.
[25] L. L. LITTLEJOHN AND A. M. KRALL, *Orthogonal polynomials and singular Sturm-Liouville systems*, I, Rocky Mt. J. Math., 16 (1986), pp. 435-479.
[26] ———, *Orthogonal polynomials and singular Sturm-Liouville systems*, II, submitted.
[27] H. D. NIESSEN, *Zum verallgemeinerten zweiten Weylschen satz*, Arch. Math., 22 (1971), pp. 648-656.
[28] ———, *Singulare S-hermitesche Rand-Eigenwertprobleme*, Manuscripta Math., 3 (1970), pp. 35-68.
[29] ———, *Greensche Matrix und die Formel von Titchmarsh-Kodaira fur singulare S-hermitesche Eigenwertprobleme*, J. Reine. Angew. Math., 261 (1972), pp. 164-193.
[30] E. C. TITCHMARSH, *Eigenfunction Expansions*, Oxford University Press, Oxford, 1962.
[31] P. W. WALKER, *A vector-matrix formulation for formally symmetric ordinary differential equations with applications to solutions of integrable square*, J. London Math. Soc., 9 (1974), pp. 151-159.
[32] H. WEYL, *Uber gewöhnliche Differentialglichungen mit Singularitaten und die zugehörigen Entwicklungen Willkurlicher Functionen*, Math. Ann., 68 (1910), pp. 220-269.
[33] J. B. MCLEOD, *The number of integrable-square solutions of ordinary differential equations*, Quart. J. Math. Oxford, 17 (1966), pp. 285-290.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 701-715, May 1989

013

$M(\lambda)$ THEORY FOR SINGULAR HAMILTONIAN SYSTEMS WITH TWO SINGULAR POINTS*

ALLAN M. KRALL†

Abstract. The $2n$-dimensional Hamiltonian system $JY' = (\lambda A + B)Y$ on an interval (a, b) is considered, where both a and b are singular points. A Green's function is derived using separated singular boundary conditions, and it is used to show that the singular boundary value problem consisting of the differential equation and boundary conditions is self-adjoint. Then a doubly singular version of Green's formula is derived and all self-adjoint boundary value problems arising from the differential equation are characterized. Finally, the spectral measure, generalized Fourier transform of an arbitrary function and its inverse transform for the original boundary value problem with separated boundary conditions are derived.

Key words. single boundary value problem, Sturm-Liouville problem, spectral resolution

AMS(MOS) subject classifications. 34B05, 34B20, 34B25

1. Introduction. The present paper follows up on our earlier work [11], which discussed linear Hamiltonian boundary value problems with one singular point. Our purpose here is to state the results for doubly singular problems and to give proofs where necessary. But where verifications closely parallel the single singular point development, we will merely refer to the proofs given in [11].

We refer to [11] or to [4] for a lengthy development of the history of the problem. We would be remiss, however, if we did not mention those works that had the most direct impact on what follows. Kodaira (see [3]) and Coddington and Levinson [3], [10] developed the theory of singular scalar differential operators of nth order. Many further details, as well as other results, were worked out by Everitt (see [11]). The notation and general techniques for systems were developed by Atkinson [1]. Following Atkinson's lead, Hinton and Shaw [6]-[8] made major advances. The present work uses the results of these authors extensively throughout. It is virtually impossible to give full credit to them where warranted, because their influence is so widespread. We are sure, however, that those who read on will recognize where their contributions appear. Without them this current version could not have been written.

We follow the footsteps of Atkinson [1] and Hinton and Shaw [6]-[8] by considering the $2n$-dimensional system $JY' = (\lambda A + B)Y$ over an interval (a, b) where both a and b are singular points. Our procedure is to first consider a problem on $[a', b']$ where $a < a' < b' < b$, then permit a' to approach a, b' to approach b.

The first problem we encounter is the verification of the existence of solutions that are in $L_A^2(a, c)$ and $L_A^2(c, b)$, $a < c < b$, where the Hilbert spaces L_A^2 are generated by

$$\langle Y, Z\rangle = \int Z^* A Y \, dt.$$

For $\operatorname{Im}\lambda \neq 0$ it can be shown that there exist at least n L_A^2-solutions toward a and toward b. By using these solutions appropriately, we can construct a Green's function

* Received by the editors May 18, 1987; accepted for publication (in revised form) August 9, 1988. This work was 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.

† Department of Mathematics, McAllister Building, Pennsylvania State University, University Park, Pennsylvania 16802.

with which we can develop a second problem to determine the self-adjointness of a differential operator related to $JY' - BY$ in $L_A^2(a, b)$.

Using the Green's function appropriately enables us to develop a singular Green's formula. This in turn permits us to determine other problems, possibly with mixed boundary conditions, that are self-adjoint in $L_A^2(a, b)$. Finally, employing the regular eigenfunction expansion over an interval $[a', b']$, we follow the lead of Coddington and Levinson [3] to develop the spectral resolution for a doubly singular problem with separated boundary conditions. Brauer [2] has developed the broader theory for mixed boundary problems.

2. Notation and definitions. We consider over an interval (a, b) the differential expression

$$(*) \qquad JY' = (\lambda A + B)Y,$$

where Y is of dimension $2n$ (a $2n \times 1$ matrix),

$$J = \begin{Bmatrix} 0 & -I_n \\ I_n & 0 \end{Bmatrix}, \qquad A = \begin{Bmatrix} A_{11}(x) & A_{12}(x) \\ A_{21}(x) & A_{22}(x) \end{Bmatrix}, \qquad B = \begin{Bmatrix} B_{11}(x) & B_{12}(x) \\ B_{21}(x) & B_{22}(x) \end{Bmatrix}$$

are locally integrable $2n \times 2n$ matrices, $A = A^* \geqq 0$, $B = B^*$. We will assume that both a and b are singular points; that is, either or both a and b may be infinite. Neither A nor B is necessarily integrable in a neighborhood of a or b. Our setting is $L_A^2(a, b)$, the Hilbert space generated by the inner product

$$\langle Y, Z \rangle = \int_a^b Z^* A Y \, dt.$$

To ensure that elements of the maximal operator are dense in $L_A^2(a, b)$, we assume that if $JY' - BY = AF$ and $AY = 0$, then $Y = 0$.

As a preliminary step we consider $(*)$ over a subinterval $[a', b']$, $a < a' < b' < b$, and impose at a' and b' the separated regular self-adjoint boundary conditions

$$(\alpha_1, \alpha_2) Y(a') = 0, \qquad (\beta_1, \beta_2) Y(b') = 0,$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are $n \times n$ matrices satisfying

$$\begin{aligned} &\operatorname{rank}(\alpha_1, \alpha_2) = n, && \operatorname{rank}(\beta_1, \beta_2) = n, \\ &\alpha_1 \alpha_1^* + \alpha_2 \alpha_2^* = I_n, && \beta_1 \beta_1^* + \beta_2 \beta_2^* = I_n, \\ &\alpha_1 \alpha_2^* - \alpha_2 \alpha_1^* = 0, && \beta_1 \beta_2^* - \beta_2 \beta_1^* = 0. \end{aligned}$$

Equation $(*)$, together with these boundary conditions, defines a regular self-adjoint boundary value problem over $[a', b']$.

Finally, let c be in $[a', b']$, and let $\mathcal{Y}(x, \lambda)$ be a fundamental matrix for $(*)$ satisfying $\mathcal{Y}(c, \lambda) = I_{2n}$. We decompose into $2n \times n$ matrices θ and ϕ such that $\mathcal{Y}(x, \lambda) = (\theta(x, \lambda), \phi(x, \lambda))$.

3. $M(\lambda)$ functions, limit circles, L^2 solutions. If $\operatorname{Im} \lambda \neq 0$, we attempt to satisfy the b' boundary condition by $\chi_{b'}(x, \lambda) = \theta(x, \lambda) + \phi(x, \lambda) M_{b'}(\lambda)$. Insertion into $(\beta_1, \beta_2) Y(b') = 0$ shows that

$$M_{b'}(\lambda) = -[(\beta_1, \beta_2)\phi(b', \lambda)]^{-1}[(\beta_1, \beta_2)\theta(b', \lambda)].$$

The inverse must exist, for otherwise the boundary value problem of $(*)$, the b' boundary condition, and the c boundary condition $(I_n, 0) Y(c) = 0$ would be self-adjoint, but would have a complex eigenvalue.

The circle equation, satisfied by $M_{b'}$, is $\pm\chi_{b'}(b',\lambda)^*(J/i)\chi_{b'}(b',\lambda)=0$. If at b', we let

$$\begin{pmatrix}\mathscr{A} & \mathscr{B}^*\\ \mathscr{B} & \mathscr{D}\end{pmatrix}=\begin{matrix}\mathscr{Y}^*(J/i)\mathscr{Y}, & \operatorname{Im}\lambda>0,\\ -\mathscr{Y}^*(J/i)\mathscr{Y}, & \operatorname{Im}\lambda<0,\end{matrix}$$

the circle equation can be written as

$$(I_n, M_{b'}^*)\begin{pmatrix}\mathscr{A} & \mathscr{B}^*\\ \mathscr{B} & \mathscr{D}\end{pmatrix}\begin{pmatrix}I_n\\ M_{b'}\end{pmatrix}=0.$$

Expanded, this is

$$M_{b'}^*\mathscr{D}M_{b'}+M_{b'}^*\mathscr{B}+\mathscr{B}^*M_{b'}+\mathscr{A}=0.$$

It is possible to show that

$$\mathscr{A}=\pm\theta^*(b')(J/i)\theta(b'),$$

$$\mathscr{B}=\pm\phi^*(b')(J/i)\theta(b')=\pm\left[2\operatorname{Im}\lambda\int_c^{b'}\phi^*A\theta\,dt-iI_n\right],$$

$$\mathscr{D}=\pm\phi^*(b')(J/i)\phi(b')=\pm\left[2\operatorname{Im}\lambda\int_c^{b'}\phi^*A\phi\,dt\right].$$

If $R_1=\mathscr{D}^{-1/2}$, $R_2=R_1(\bar\lambda)$, and $C=-\mathscr{D}^{-1}\mathscr{B}$, then $M_{b'}=C+R_1U_bR_2$, where U_b is any unitary matrix. As b' approaches b, C, R_1, and R_2 have limits C_b, R_b, $\tilde R_b$, giving M_b.

It is shown in [11] that as b' approaches b, $M_{b'}$ can be made to approach $M_b=C_b+R_bU_b\tilde R_b$, where

$$C_b=\lim_{b'\to b}\left[2\operatorname{Im}\lambda\int_c^{b'}\phi^*A\phi\,dt\right]^{-1}\left[2\operatorname{Im}\lambda\int_c^{b'}\phi^*A\theta\,dt-iI_n\right],$$

$$R_b=\lim_{b'\to b}\left[2\,|\operatorname{Im}\lambda|\int_c^{b}\phi^*A\phi\,dt\right]^{-1/2},\qquad \tilde R_b(\lambda)=R_b(\bar\lambda),$$

and $U_b(\lambda)$ is a unitary matrix.

It is further shown that if $\chi_b(x,\lambda)=\theta(x,\lambda)+\phi(x,\lambda)M_b(\lambda)$, then

$$\int_c^b\chi_b^*A\chi_b\,dt\leqq[M_b-M_b^*]/2\operatorname{Im}\lambda.$$

For $\chi_b(x,\lambda)$ and $\chi_b(x,\mu)$, $\operatorname{Im}\lambda\neq0$, $\operatorname{Im}\mu\neq0$, we have

$$\lim_{x\to b}\chi_b(x,\mu)^*J\chi_b(x,\lambda)=0.$$

Note that for all λ, $\operatorname{Im}\lambda\neq0$,

$$\frac{\operatorname{Im}M_b}{\operatorname{Im}\lambda}=\frac{M_b-M_b^*}{2i\operatorname{Im}\lambda}>0.$$

Similarly, we attempt to satisfy the a' boundary conditions by $\chi_{a'}(x,\lambda)=\theta(x,\lambda)+\phi(x,\lambda)M_{a'}(\lambda)$. Insertion into $(\alpha_1,\alpha_2)Y(a')=0$ shows that

$$M_{a'}=-[(\alpha_1,\alpha_2)\phi(a')]^{-1}[(\alpha_1,\alpha_2)\theta(a')],$$

where, again, the inverse must exist.

The circle equation, satisfied by $M_{a'}$, is $\pm\chi_{a'}(a',\lambda)^*(J/i)\chi_{a'}(a',\lambda)=0$. If at a', we let

$$\begin{pmatrix}\mathscr{A} & \mathscr{B}^*\\ \mathscr{B} & \mathscr{D}\end{pmatrix}=\begin{matrix}-\mathscr{Y}^*(J/i)\mathscr{Y}, & \operatorname{Im}\lambda>0,\\ \mathscr{Y}^*(J/i)\mathscr{Y}, & \operatorname{Im}\lambda<0,\end{matrix}$$

the circle equation can be rewritten as

$$(I_n, M_{a'}^*)\begin{pmatrix}\mathscr{A} & \mathscr{B}\\ \mathscr{B} & \mathscr{D}\end{pmatrix}\begin{pmatrix}I_n\\ M_{a'}\end{pmatrix}=0.$$

Expanded, this is again

$$M_{a'}^*\mathscr{D}M_{a'}+M_{a'}^*\mathscr{B}+\mathscr{B}^*M_{a'}+\mathscr{A}=0.$$

It is possible to show that

$$\mathscr{A}=\pm\theta^*(a')(J/i)\theta(a'),$$

$$\mathscr{B}=\pm\phi^*(a')(J(i)\theta(a'))=\pm\left[2\operatorname{Im}\lambda\int_{a'}^{c}\phi^*A\theta\,dt+iI_n\right],$$

$$\mathscr{D}=\pm\phi^*(a')(J(i)\phi(a'))=\pm\left[2\operatorname{Im}\lambda\int_{a'}^{c}\phi^*A\phi\,dt\right].$$

If $R_1=\mathscr{D}^{1/2}$, $R_2=R_1(\bar\lambda)$ and $C=-\mathscr{D}^{-1}\mathscr{B}$, then $M_{a'}=C+R_1U_aR_2$, where U_a is any unitary matrix. As a' approaches a, C, R_1 and R_2 have limits C_a, R_a, $\tilde R_a$, giving M_a.

As a' approaches a, $M_{a'}$ can be made to approach $M_a=C_a+R_aU_a\tilde R_a$ where

$$C_a=\lim_{a'\to a}-\left[2\operatorname{Im}\lambda\int_a^c\phi^*A\phi\,dt\right]^{-1}\left[2\operatorname{Im}\lambda\int_{a'}^{c}\phi^*A\theta\,dt+iI_n\right],$$

$$R_a=\lim_{a'\to a}\left[2|\operatorname{Im}\lambda|\int_{a'}^{c}\phi^*A\phi\,dt\right]^{-1/2},\qquad \tilde R_a=R_a(\bar\lambda),$$

and U_a is a perhaps different, unitary matrix.

With $\chi_a(x,\lambda)=\theta(x,\lambda)+\phi(x,\lambda)M_a(\lambda)$, it is also true that

$$\int_a^c\chi_a^*A\chi_a\,dt\leqq[M_a^*-M_a]/2i\operatorname{Im}\lambda.$$

For $\chi_a(x,\lambda)$ and $\chi_a(x,\mu)$, $\operatorname{Im}\lambda\neq0$, $\operatorname{Im}\mu\neq0$, we have

$$\lim_{x\to a}\chi_a(x,\mu)^*J\chi_a(x,\lambda)=0.$$

Note that

$$\frac{\operatorname{Im}M_a}{\operatorname{Im}\lambda}=\frac{M_a-M_a^*}{2i\operatorname{Im}\lambda}<0.$$

We will require some facts concerning $M_a(\lambda)$ and $M_b(\lambda)$.

THEOREM 3.1. *For all* λ, $\operatorname{Im}\lambda\neq0$,

(a) $M_a(\lambda)\neq M_b(\lambda)$, $[\operatorname{Im}\lambda][M_a(\lambda)-M_b(\lambda)]<0$;

(b) $M_a(\lambda)=M_a(\bar\lambda)^*$, $M_b(\lambda)=M_b(\bar\lambda)^*$;

(c) $M_b(\lambda)$, $M_b(\lambda)$, $M_a(\lambda)-M_b(\lambda)$ *are all invertible*;

(d) $M_0(\lambda)[M_a(\lambda)-M_b(\lambda)]^{-1}M_a(\lambda)=M_a(\lambda)[M_a(\lambda)-M_b(\lambda)]^{-1}M_b(\lambda)$.

Proof. The second part follows from the statements

$$\lim_{x\to a}\chi_a(x,\mu)^*J\chi_a(x,\lambda)=0,$$

$$\lim_{x\to b}\chi_b(x,\mu)^*J\chi_b(x,\lambda)=0,$$

letting $\mu = \bar{\lambda}$. The third follows from noting that for any matrices $M = A + iB$, with $M^* = A - iB$, we have $A = (M + M^*)/2$ and $B = (M - M)^*/2i$. If $B > 0$ or $B < 0$, then M is nonsingular. Suppose M is singular. Then there exists an eigenvector v such that $Mv = 0$. This implies $0 = v^*Mv = v^*Av + iv^*Bv$. Since $B > 0$ or $B < 0$, iv^*Bv is imaginary, while v^*Av is real. This is impossible.

The fourth is an easy computation.

4. The differential operator. From an extension of a theorem of Everitt [5], we know that the number of solutions of (*) in $L^2_A(a, b)$ is invariant provided $\operatorname{Im} \lambda > 0$ or $\operatorname{Im} \lambda < 0$. It is possible, however, for the deficiency indices to be unequal. Only when both ends a and b are limit circle (all solutions in $L^2_A(a, b)$ for all λ), or when A and B are real, are they guaranteed to be equal. Here we must assume the deficiency indices are equal.

DEFINITION 4.1. We denote by D_M those elements Y in $L^2_A(a, b)$ satisfying

$$lY = JY' - BY = AF \tag{1}$$

exists almost everywhere for some F in $L^2_A(a, b)$.

DEFINITION 4.2. We define the maximal operator L_M by setting $L_M Y = F$ for all Y in D_M.

THEOREM 4.3. *Let Y_j be a solution of*

$$JY_j' = (\bar{\lambda}A + B)Y_j, \qquad \operatorname{Im} \lambda \neq 0.$$

Then for all Y in D_M,

$$B_{a_j}(Y) = \lim_{x \to a} Y_j^* JY$$

exists if and only if Y_j is in $L^2_A(a, c)$,

$$B_{b_j}(Y) = \lim_{x \to b} Y_j^* JY$$

exists if and only if Y_j is in $L^2_A(c, b)$.

DEFINITION 4.4. Let $\operatorname{Im} \lambda \neq 0$. Let $M_a(\bar{\lambda}) = \tilde{C}_a + \tilde{R}_a U_a R_a$ be on the limit circle at a. Let $\chi_a(x, \bar{\lambda}) = \theta(x, \bar{\lambda}) + \phi(x, \bar{\lambda}) M_a(\bar{\lambda})$ satisfy (*) with λ replaced by $\bar{\lambda}$ and be in $L^2_A(a, c)$. Let $M_b(\bar{\lambda}) = \tilde{C}_b + \tilde{R}_b U_b R_b$ be on the limit circle at b. Let $\chi_b(x, \bar{\lambda}) = \theta(x, \bar{\lambda}) + \phi(x, \bar{\lambda}) M_b(\bar{\lambda})$ satisfy (*) with λ replaced by $\bar{\lambda}$ and be in $L^2_A(c, b)$. We define the boundary values $B_a(Y)$ and $B_b(Y)$ by setting

$$B_a(Y) = \lim_{x \to a} \chi_a(x, \bar{\lambda})^* JY(x),$$

$$B_b(Y) = \lim_{x \to b} \chi_b(x, \bar{\lambda})^* JY(x),$$

for all Y in D_M.

Note that $\bar{\lambda}$ is used in the definition. This is for convenience only and will be removed later.

DEFINITION 4.5. We denote by D those elements Y in $L^2_A(a, b)$ satisfying (1) $l(Y) = JY' - BY = AF$ exists almost everywhere for some F in $L^2_A(a, b)$; (2) $B_a(Y) = 0$; (3) $B_b(Y) = 0$, for some fixed λ, $\operatorname{Im} \lambda \neq 0$.

DEFINITION 4.6. We define the operator L by setting $LY = F$ for all Y in D.

5. The resolvent, the Green's function. The inverse of $(L - \lambda I)$ can be calculated with ease. We solve (*) together with the two boundary conditions. If we set $Y = \mathcal{Y}C$, variation of parameters shows $C' = -J\mathcal{Y}^* AF$, where $\mathcal{Y}'(x, \lambda) = \mathcal{Y}(x, \bar{\lambda})^*$. Thus

$$Y(x) = -\mathcal{Y}(x, \lambda) \int_c^x J\mathcal{Y}'(\xi, \lambda) A(\xi) F(\xi)\, d\xi + \mathcal{Y}(x, \lambda) K.$$

We multiply by

$$\begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} J,$$

so

$$\begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} JY(x) = -\begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} JY \int_c^x JY^t AF\, d\xi + \begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} JYK.$$

Now

$$\begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} JY$$

is constant. At $x = c$,

$$\begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} JY = \begin{pmatrix} M_b^t & -I_n \\ 0 & 0 \end{pmatrix}.$$

Thus

$$\begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} JY(x) = \int_c^x \begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} AF\, d\xi + \begin{pmatrix} M_b^t & -I_n \\ 0 & 0 \end{pmatrix} K.$$

Letting $x \to b$, we get

$$0 = \int_c^b \begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} AF\, d\xi + \begin{pmatrix} M_b & -I_n \\ 0 & 0 \end{pmatrix} K.$$

Note that the integral exists since χ_b is in $L_A^2(c, b)$.

Now return to Y and multiply by

$$\begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} J.$$

Then

$$\begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} JY(x) = \begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} J\mathcal{Y} \int_x^c J\mathcal{Y}^t AF\, d\xi + \begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} J\mathcal{Y}K.$$

Now

$$\begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} J\mathcal{Y}$$

is constant. At $x = c$

$$\begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} J\mathcal{Y} = \begin{pmatrix} 0 & 0 \\ M_a^t & -I_n \end{pmatrix}.$$

Thus

$$\begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} JY(x) = -\int_x^c \begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} AF\, d\xi + \begin{pmatrix} 0 & 0 \\ M_a^t & -I_n \end{pmatrix} K.$$

Letting $x \to a$, we get

$$0 = -\int_a^c \begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} AF\, d\xi + \begin{pmatrix} 0 & 0 \\ M_a^t & -I_n \end{pmatrix} K.$$

Again the integral exists since χ_a is in $L_A^2(a, c)$.

Add the two limit equations together to get

$$0=\int_c^b \begin{pmatrix} \chi_b^t \\ 0 \end{pmatrix} AF\,d\xi - \int_a^c \begin{pmatrix} 0 \\ \chi_a^t \end{pmatrix} AF\,d\xi + \begin{bmatrix} M_b^t & -I_n \\ M_a^t & -I_n \end{bmatrix} K.$$

The coefficient of K is nonsingular, yielding

$$K=\begin{bmatrix} (M_a-M_b)^{-1} & -(M_a-M_b)^{-1} \\ M_a(M_a-M_b)^{-1} & -M_b(M_a-M_b)^{-1} \end{bmatrix} \int_c^{b'} \begin{bmatrix} I_n & M_b^t \\ 0 & 0 \end{bmatrix} \mathcal{Y}^t AF\,d\xi$$
$$-\begin{bmatrix} (M_a-M_b)^{-1} & -(M_a-M_b)^{-1} \\ M_a(M_a-M_b)^{-1} & -M_b(M_a-M_b)^{-1} \end{bmatrix} \int_a^c \begin{pmatrix} 0 & 0 \\ I_n & M_a^t \end{pmatrix} \mathcal{Y}^t AF\,d\xi.$$

Inserting this in Y, we obtain

$$Y(x)=\mathcal{Y}\begin{pmatrix} I_n & 0 \\ M_b & 0 \end{pmatrix}\begin{pmatrix} 0 & (M_a-M_b)^{-1} \\ (M_a-M_b)^{-1} & 0 \end{pmatrix} \int_a^x \begin{pmatrix} 0 & 0 \\ I_n & M_a^t \end{pmatrix} \mathcal{Y}^t AF\,d\xi$$
$$+\mathcal{Y}\begin{pmatrix} 0 & I_n \\ 0 & M_a \end{pmatrix}\begin{pmatrix} 0 & (M_a-M_b)^{-1} \\ (M_a-M_b)^{-1} & 0 \end{pmatrix} \int_x^b \begin{pmatrix} I_n & M_b^t \\ 0 & 0 \end{pmatrix} \mathcal{Y}^t AF\,d\xi.$$

Further computation shows this can be written as

$$Y(x)=\chi_b(x,\lambda)(M_a(\lambda)-M_b(\lambda))^{-1}\int_a^x \chi_a(\xi,\bar{\lambda})^* A(\xi)F(\xi)\,d\xi$$
$$+\chi_a(x,\lambda)(M_a(\lambda)-M_b(\lambda))^{-1}\int_x^b \chi_b(\xi,\bar{\lambda})^* A(\xi)F(\xi)\,d\xi.$$

THEOREM 5.1. *$(L-\lambda I)^{-1}$ is given by*

$$(L-\lambda I)^{-1}F(x)=\int_a^b G(\lambda,x,\xi)A(\xi)F(\xi)\,d\xi,$$

where

$$G(\lambda,x,\xi)=\chi_b(x,\lambda)(M_a(\lambda)-M_b(\lambda))^{-1}\chi_a(\xi,\bar{\lambda})^*, \qquad a<\xi<x<b,$$
$$G(\lambda,x,\xi)=\chi_a(x,\lambda)(M_a(\lambda)-M_b(\lambda))^{-1}\chi_b(\xi,\bar{\lambda})^*, \qquad a<x<\xi<b.$$

THEOREM 5.2. *G is symmetric*

$$G(\lambda,x,\xi)=G(\bar{\lambda},\xi,x)^*.$$

THEOREM 5.3. *$(L-\lambda I)^{-1}$ is bounded*

$$\|(L-\lambda I)^{-1}\|\leqq 1/|\operatorname{Im}\lambda|.$$

THEOREM 5.4. *L is self-adjoint.*

6. Parameter independence of the domain. It appears that D is dependent on the parameter λ used in the boundary conditions. Indeed, the Green's function was only calculated for λ. As shown in [11], however, we can demonstrate that λ may vary with impunity so long as $\operatorname{Im}\lambda\neq 0$.

THEOREM 6.1. *Let Y be in D. Then for all λ, $\operatorname{Im}\lambda\neq 0$, $B_a(Y)=0$, $B_b(Y)=0$.*

THEOREM 6.2. *For all λ, $\operatorname{Im}\lambda\neq 0$, $(L-\lambda I)^{-1}$ exists and is given by* (**)

$$\|(L-\lambda)^{-1}\|\leqq 1/|\operatorname{Im}\lambda|.$$

7. Green's formula. Only minor modifications in the technique used in [11] are needed to express the Lagrange bilinear form Z^*JY properly for insertion into Green's formula. Using the Green's function for the doubly singular problem, we get

$$G(\lambda, x, \xi) = \chi_b(x, \lambda)[M_a(\lambda) - M_b(\lambda)]^{-1}\chi_a(\xi, \bar{\lambda})^*, \qquad a \leqq \xi < x \leqq b,$$

$$G(\lambda, x, \xi) = \chi_a(x, \lambda)[M_a(\lambda) - M_b(\lambda)]^{-1}\chi_b(\xi, \bar{\lambda})^*, \qquad a \leqq x < \xi \leqq b;$$

instead of the one for the previous problem [11]. We set

$$R_Y(x) = \int_a^b G(\lambda x, \xi)A(\xi)F(\xi)\, d\xi.$$

The previous calculation [11] then goes through without change.

Assume that for $\operatorname{Im}\lambda \neq 0$ there are m solutions of $(*)$ in $L_A^2(c, b)$, $a < c < b$. Since n of these are represented by χ_b, the other $m-n$ are linear combinations of columns of ϕ.

THEOREM 7.1. *Let $\phi(x, \lambda) = (\phi_1, \phi_2)(x, \lambda)E_b$, where ϕ_1 consists of all of the $\phi - L_A^2(c, b)$ solutions of $(*)$. Let $\phi(x, \bar{\lambda}) = (\phi_1, \phi_2)(x, \bar{\lambda})\tilde{E}_b$, where ϕ_1 consists of all of the $\phi - L_A^2(c, b)$ solutions of $(*)$ with λ replaced by $\bar{\lambda}$. Let $(\chi_1, \chi_2)(x, \bar{\lambda}) = \chi_b(x, \bar{\lambda})E_b^*$ and let $(\chi_1, \chi_2)(x, \lambda) = \chi_b(x, \lambda)\tilde{E}_b^*$. Then for all Y and Z in D_M,*

$$\lim_{x\to b} \chi_2(x, \bar{\lambda})^*JY(x) = 0, \qquad \lim_{x\to b} \chi_2(x, \lambda)^*JZ(x) = 0.$$

Returning to $\lim_{x\to b} Z(x)^*JY(x)$, we find

$$\lim_{x\to b} Z(x)^*JY(x) = [\lim_{x\to b} \phi_1(x, \lambda)^*JZ(x)]^*[\lim_{x\to b} \chi_1(x, \bar{\lambda})^*JY(x)]$$

$$-[\lim_{x\to b} \chi_1(x, \lambda)^*JZ(x)]^*[\lim_{x\to b} \phi_1(x, \bar{\lambda})^*JY(x)].$$

Hence we formulate the following theorem.

THEOREM 7.2. *Under the conditions of Theorem* 7.1, *let*

$$B_b(Y) = \begin{pmatrix} \lim_{x\to b} \chi_1(x, \bar{\lambda})^*JY(x) \\ \lim_{x\to b} \phi_1(x, \bar{\lambda})^*JY(x) \end{pmatrix},$$

$$\tilde{B}_b(Z) = \begin{pmatrix} \lim_{x\to b} \chi_1(x, \lambda)^*JZ(x), \\ \lim_{x\to b} \phi_1(x, \lambda)^*JZ(x) \end{pmatrix},$$

$$J_b = \begin{pmatrix} 0 & -I_{m-n} \\ I_{m-n} & 0 \end{pmatrix},$$

where (m, m) are the defect indices of $(*)$ at b. Then

$$\lim_{x\to b} Z(x)^*JY(x) = \tilde{B}_b(Z)^*J_bB_b(Y).$$

The symbol $\tilde{}$ indicates λ is used instead of $\bar{\lambda}$.

Assume that for $\operatorname{Im}\lambda \neq 0$, there are p solutions of $(*)$ in $L_A^2(a, c)$, $a < c < b$. Since n of these are represented by χ_a, the other $p-n$ are linear combinations of ϕ.

THEOREM 7.3. *Let* $\phi(x,\lambda)=(\phi_1,\phi_2)(x,\lambda)E_a$, *where* ϕ_1 *consists of all of the* $\phi - L^2_A(a,c)$ *solutions of* $(*)$. *Let* $\phi(x,\bar\lambda)=(\phi_1,\phi_2)(x,\lambda)\tilde E_a$, *where* ϕ_1 *consists of all the* $\phi - L^2_A(a,c)$ *solutions of* $(*)$ *with* λ *replaced by* $\bar\lambda$. *Let* $(\chi_1,\chi_2)(x,\bar\lambda)=\chi_a(x,\bar\lambda)E_a^*$, *and let* $(\chi_1,\chi_2)(x,\lambda)=\chi_a(x,\lambda)\tilde E_a^*$. *Then for all* Y *and* Z *in* D_M,

$$\lim_{x\to a}\chi_2(x,\bar\lambda)^*JY(x)=0, \qquad \lim_{x\to a}\chi_2(x,\lambda)^*JZ(x)=0.$$

THEOREM 7.4. *Under the conditions of Theorem* 7.3, *let*

$$B_a(Y)=\begin{pmatrix}\lim_{x\to a}\chi_1(x,\bar\lambda)^*JY(x)\\ \lim_{x\to a}\phi_1(x,\bar\lambda)^*JY(x)\end{pmatrix},$$

$$\tilde B_a(Z)=\begin{pmatrix}\lim_{x\to a}\chi_1(x,\lambda)^*JZ(x)\\ \lim_{x\to a}\phi_1(x,\lambda)^*JZ(x)\end{pmatrix},$$

$$J_a=\begin{pmatrix}0 & -I_{p-n}\\ I_{p-n} & 0\end{pmatrix},$$

where (p,p) *are the defect indices of* $(*)$ *at* a; *then*

$$\lim_{x\to a}Z^*(x)JY(x)=\tilde B_a(Z)^*J_aB_a^*(Y).$$

Again the symbol $\tilde{}$ indicates λ is used instead of $\bar\lambda$.

The pieces now can be put together to yield

$$\int_a^b\{Z^*[JY'-BY]-[JZ'-BZ]^*Y\}\,dx=(\tilde B_a(Z)^*,\tilde B_b(Z)^*)\begin{pmatrix}-J_a & 0\\ 0 & J_b\end{pmatrix}\begin{pmatrix}B_a(Y)\\ B_b(Y)\end{pmatrix}.$$

If M, N, P, Q are $r\times(2p-2n)$, $r\times(2m-2n)$, $(2m+2p-4n-r)\times(2p-2n)$, and $(2m+2p-4n-r)\times(2m-2n)$ matrices, $0\leqq r\leqq 2m+2p$, and $\tilde M$, $\tilde N$, $\tilde P$, $\tilde Q$ are chosen so that

$$\begin{pmatrix}\tilde M^* & \tilde P^*\\ \tilde N^* & \tilde Q^*\end{pmatrix}\begin{pmatrix}M & N\\ P & Q\end{pmatrix}=\begin{pmatrix}-J_a & 0\\ 0 & J_b\end{pmatrix}.$$

We finally have the following theorem.

THEOREM 7.5 (Green's formula). *Let* Y *and* Z *be in* D_M. *Then*

$$\begin{aligned}\int_a^b&\{Z^*[JY'-BY]-[JZ'-BZ]^*Y\}\,dx\\ &=[\tilde M\tilde B_a(Z)+\tilde NB_b(Z)]^*[MB_a(Y)+NB_b(Y)]\\ &\quad+[\tilde P\tilde B_a(Z)+\tilde QB_b(Z)]^*[PB_a(Y)+QB_b(Y)].\end{aligned}$$

DEFINITION 7.6. We denote by $\tilde D$ those elements Y in $L^2_A(a,b)$ satisfying

(1) Y is in D_M;

(2) $MB_a(Y)+NB_b(Y)=0$.

DEFINITION 7.7. We denote by $\tilde L$ the operator defined by setting $\tilde LY=F$ whenever $JY'-BY=AF$ and Y is in $\tilde D$.

DEFINITION 7.8. We denote by $\tilde D^*$ those elements Z in $L^2_A(a,b)$ satisfying

(1) Z is in D_M;

(2) $\tilde P\tilde B_a(Z)+\tilde Q\tilde B_b(Z)=0$.

DEFINITION 7.9. We denote by $\tilde L^*$ the operator defined by setting $\tilde L^*Z=G$ whenever $JZ'-BZ=AG$ and Z is in $\tilde D^*$.

THEOREM 7.10. *The abuse of notation above is correct. The adjoint of* $\tilde{L}$ *in* $L_A^2(a, b)$ *is* $\tilde{L}^*$. *The adjoint of* $\tilde{L}^*$ *in* $L_A^2(a, b)$ *is* $\tilde{L}$.

The proof is almost the same as that with only one singular point.

8. Self-adjoint problems. In [11] it was shown that every boundary condition at b can be written as a linear combination of terms from

$$B_b(Y) = \begin{bmatrix} \lim_{x\to b} \chi_1(x, \bar{\lambda})^* JY(x) \\ \lim_{x\to b} \phi_1(x, \bar{\lambda})^* JY(x) \end{bmatrix}.$$

In particular $\tilde{B}_b(Y)$ has the representation

$$B_b(Y) = V\tilde{B}_b(Y),$$

where

$$V_b = \begin{pmatrix} V_{11} & 0 \\ V_{21} & V_{22} \end{pmatrix}$$

is nonsingular. If λ can be made real, then $V_b = I$.

This same sort of representation also holds at a:

$$B_a(Y) = V_a\tilde{B}_a(Y).$$

As an immediate consequence, we find the following theorem holds.

THEOREM 8.1. $\tilde{L}$ *is self-adjoint if and only if* $r = m + p - 2n$ *and*

$$MV_aJ_aM^* = NV_bJ_bN^*.$$

9. The spectral resolution of L. The operator L was defined in § 4. Restricted to a finite regular subinterval (a', b'), the resulting operator has a well-known spectral resolution that is an eigenfunction expansions: if $\{\lambda_k\}_{k=1}^{\infty}$ are the eigenvalues with $\chi_k = \theta(x, \lambda_k)S_k + \phi(x, \lambda_k)T_k$, $k = 1, \cdots$, the associated normalized eigenfunctions, then

$$\chi_k = \mathcal{Y}(x, \lambda_k)\begin{pmatrix} S_k \\ T_k \end{pmatrix},$$

and we have the following theorem.

THEOREM 9.1. *For all* F *in* $L_A^2(a', b')$,

$$F(x) = \sum_{k=1}^{\infty} \mathcal{Y}(x, \lambda_k)\begin{pmatrix} S_k \\ T_k \end{pmatrix}(S_k^*, T_k^*)\int_{a'}^{b'} \mathcal{Y}(\xi, \lambda_k)^* A(\xi)F(\xi)\, d\xi.$$

Parseval's equality also holds.

THEOREM 9.2. *For all* F *in* $L_A^2(a', b')$,

$$\int_{a'}^{b'} F^*AF\, d\xi = \left[\int_{a'}^{b'} \mathcal{Y}(\xi, \lambda_k)^* AF\, d\xi\right]^* \begin{pmatrix} S_kS_k^* & S_kT_k^* \\ T_kS_k^* & T_kT_k^* \end{pmatrix}\left[\int_{a'}^{b'} \mathcal{Y}(\xi, \lambda_k)^* AF\, d\xi\right].$$

DEFINITION 9.3. Let R_k^2 denote the $2n \times 2n$ matrix

$$\begin{pmatrix} S_kS_k^* & S_kT_k^* \\ T_kS_k^* & T_kT_k^* \end{pmatrix}.$$

DEFINITION 9.4. Let

$$G_I(\lambda) = \int_{a'}^{b'} \mathcal{Y}(\xi, \lambda)^* AF\, d\xi,$$

$$= \int_{a'}^{b'} \begin{pmatrix} \theta(\xi, \lambda)^* \\ \phi(\xi, \lambda)^* \end{pmatrix} AF\, d\xi.$$

DEFINITION 9.5. Let $P_I(\lambda)$ be a $2n \times 2n$ matrix-valued function satisfying

(1) $P_I(0+) = 0$;

(2) $P_I(\lambda)$ is increasing, jumping R_k^2 at $\lambda = \lambda_k$, but otherwise constant, continuous from above.

Thus

$$P_I(\lambda) = \sum_{0<\lambda\leqq\lambda} R_k^2, \qquad \lambda \geqq 0,$$

$$P_I(\lambda) = -\sum_{\lambda<\lambda\leqq 0} R_k^2, \qquad \lambda < 0.$$

THEOREM 9.6 (Parseval's equality). *Let F be an arbitrary element in $L_A^2(a', b')$. Then*

$$\int_{a'}^{b'} F^* A F \, d\xi = \int_{-\infty}^{\infty} G_I^*(\lambda) \, dP_I(\lambda) G_I(\lambda).$$

This can be extended by the polarization identities.

COROLLARY 9.7. *Let F_1, F_2 be arbitrary elements in $L_A^2(a', b')$. Let G_{I1}, G_{I2} correspond to them according to Definition 9.4. Then*

$$\int_{a'}^{b'} F_2^* A F_1 \, d\xi = \int_{-\infty}^{\infty} G_{12}(\lambda)^* \, dP_I(\lambda) G_{I1}(\lambda).$$

THEOREM 9.8. *There exists a nondecreasing $2n \times 2n$ matrix-valued function $P(\lambda)$, defined on $(-\infty, \infty)$ satisfying*

(1) $P(0+) = 0$;

(2) $P(\lambda) - P(\mu) = \lim_{I \to (a,b)} [P_I(\lambda) - P_I(\mu)]$, $\lambda > \mu$.

The proof consists of showing that if

$$H(x) = \chi_{b'}(x, \lambda)[M_{a'} - M_{b'}]^{-1} \chi_{a'}(c, \bar{\lambda})^*, \qquad c < x,$$

$$H(x) = \chi_{a'}(x, \lambda)[M_{a'} - M_{b'}]^{-1} \chi_{b'}(c, \bar{\lambda})^*, \qquad c > x.$$

Then $\int_{a'}^{b'} H^* A H \, dx < K$. This ultimately yields

$$\int_{-\mu}^{\mu} dP_I(\lambda) < K(1 + \mu^2).$$

Helly's first convergence theorem can then be applied. See [10].

THEOREM 9.9. *If F is in $L_A^2(a, b)$ there is a function $G(\lambda)$ in $L_p^2(-\infty, \infty)$, with inner product*

$$(G, H)_p = \int_{-\infty}^{\infty} H^* \, dP \, G,$$

such that if

$$E(\lambda) = G(\lambda) - \int_{a'}^{b'} \mathcal{Y}(\xi, \lambda)^* A(\xi) F(\xi) \, d\xi,$$

then

$$\lim_{(a',b') \to (a,b)} \int_{-\infty}^{\infty} E(\lambda)^* \, dP(\lambda) E(\lambda) = 0$$

and

$$\int_a^b F(\xi)^* A(\xi) F(\xi) \, d\xi = \int_{-\infty}^{\infty} G(\lambda)^* \, dP(\lambda) G(\lambda).$$

THEOREM 9.10. *If $G(\lambda)$ is the limit of $\int_{a'}^{b'} \mathcal{Y}(\xi,\lambda)^* A(\xi)F(\xi)\,d\xi$ in $L_p^2(-\infty,\infty)$, then*

$$\int_{-\infty}^{\infty} \mathcal{Y}(x,\lambda)\,dP(\lambda)G(\lambda) = F(x)$$

in $L_A^2(a,b)$, that is,

$$\lim_{I\to(-\infty,\infty)} \int_a^b \left[F - \int_I \mathcal{Y}\,dP\,G\right]^* A\left[F - \int_I \mathcal{Y}\,dP\,G\right] d\xi = 0.$$

Theorem 9.9 may be extended to involve inner products by use of the polarization identity. The inner product form of Parseval's equality is

$$\int_a^b F_2(\xi)^* A(\xi)F_1(\xi)\,d\xi = \int_{-\infty}^{\infty} G_2(\lambda)^*\,dP(\lambda)G_1(\lambda),$$

where

$$G_j(\lambda) = \int_a^b \mathcal{Y}(\xi,\lambda)^* A(\xi)F_j(\xi)\,d\xi, \qquad j = 1, 2.$$

Theorems 9.9 and 9.10 may be extended to represent the resolvent operator $(L-\lambda_0 I)^{-1}$ when λ_0 is not in the support of $dP(\lambda)$. Parseval's equality is

$$\int_a^b [(L-\lambda_0 I)^{-1}F(\xi)]^* A(\xi)[(L-\lambda_0 I)^{-1}F(\xi)]\,d\xi = \int_{-\infty}^{\infty} \frac{G(\lambda)^*\,dP(\lambda)G(\lambda)}{|\lambda-\lambda_0|^2}.$$

The resolvent expansion is

$$(L-\lambda_0 I)^{-1}F(x) = \int_{-\infty}^{\infty} \mathcal{Y}(x,\lambda)\,dP(\lambda)\frac{G(\lambda)}{\lambda-\lambda_0}.$$

10. The converse problem. The preceding section began with choosing an F in $L_A^2(a,b)$, producing a G in $L_p^2(-\infty,\infty)$, and then showing that F could be recovered from G. In this section we begin with G, produce F, and then recover G.

Without the assumption that $JY' - BY = 0$, $AY = 0$ implies $Y = 0$, $L_p^2(-\infty,\infty)$ may be too large in the sense that $G \to F \to \tilde{G}$, but $\tilde{G}$ may not equal G. $\tilde{G}$ may be only in a subspace of $L_p^2(-\infty,\infty)$.

With the assumption made in the Introduction (i.e., let $F = 0$), there is no such difficulty.

THEOREM 10.1. *If $G(\lambda)$ is in $L_p^2(-\infty,\infty)$, there is a unique $F(x)$ in $L_A^2(a,b)$ such that*

$$F(x) = \int_{-\infty}^{\infty} \mathcal{Y}(x,\lambda)\,dP(\lambda)G(\lambda)$$

and

$$G(\lambda) = \int_a^b \mathcal{Y}(\xi,\lambda)^* A(\xi)F(\xi)\,d\xi.$$

11. The relation between M_a, M_b, and $P(\lambda)$. The matrices M_a and M_b can frequently be determined by a careful inspection of the solutions of (*) to determine appropriate L_A^2 solutions. More difficult is the determination of the spectral matrix $P(\lambda)$, since its existence follows from Helly's selection theorems. Fortunately, they are intimately connected.

THEOREM 11.1. *Let*

$$M=\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

where

$$M_{11}=(M_a-M_b)^{-1}-(M_a^*-M_b^*)^{-1}=2i\ \mathrm{Im}\,(M_a-M_b)^{-1},$$

$$M_{12}=\tfrac{1}{2}(M_a-M_b)^{-1}(M_a+M_b)-\tfrac{1}{2}(M_a^*-M_b^*)^{-1}(M_a^*+M_b^*),$$

$$M_{21}=-\tfrac{1}{2}(M_a+M_b)(M_a-M_b)^{-1}+\tfrac{1}{2}(M_a^*+M_b^*)(M_a^*-M_b^*)^{-1}=-M_{12}^*,$$

$$M_{22}=M_a(M_a-M_b)^{-1}M_b-M_b^*(M_a^*-M_b^*)^{-1}M_a^*=2i\ \mathrm{Im}\ M_a(M_a-M_b)^{-1}M_b.$$

If λ_1 *and* λ_2 *are real, then*

$$P(\lambda_2)-P(\lambda_1)=\lim_{v\to 0+}\frac{1}{2\pi i}\int_{\lambda_1}^{\lambda_2} M(\mu+iv)\,d\mu.$$

12. The spectral resolution. We connect these results to the classic representation of the identity as an integral generated by a projection valued measure E_λ. Given

$$F(x)=\int_{-\infty}^{\infty}\mathcal{Y}(x,\lambda)\,dP(\lambda)G(\lambda),$$

where

$$G(\lambda)=\int_a^b \mathcal{Y}^*(\xi,\lambda)A(\xi)F(\xi)\,d\xi,$$

we define

$$E_\lambda F(x)=\int_{-\infty}^{\lambda+}\mathcal{Y}(x,\lambda)\,dP(\lambda)G(\lambda).$$

Considered as the limit of eigenfunction expansions, E_λ can easily be shown to be a projection that is continuous from above and satisfies $E_{\lambda_1}E_{\lambda_2}=E_{\lambda_1}$ when $\lambda_1\leqq\lambda_2$, as well as $E_{-\infty}=0$, $E_\infty=I$. If we let $\{\lambda_j\}_{j=-\infty}^{\infty}$ be a partition of $(-\infty,\infty)$, $\lambda_i<\lambda_j$ if $i<j$, and

$$\Delta_j EF(x)=\int_{\lambda_j}^{\lambda_{j+1}}\mathcal{Y}(x,\lambda)\,dP(\lambda)G(\lambda),$$

then $F(x)=\sum_{j=-\infty}^{\infty}\Delta_j EF(x)$. As $\{\lambda_j\}_{j=-\infty}^{\infty}$ becomes finer, we may write

$$F(x)=\int_{-\infty}^{\infty} dE_\lambda\,F(x)$$

as the limit of the decomposition above.

If Y is in D, it has the representation

$$Y(x)=\int_{-\infty}^{\infty}\mathcal{Y}(x,\lambda)\,dP(\lambda)G(\lambda),$$

where

$$G(\lambda)=\int_a^b \mathcal{Y}^*(\xi,\lambda)A(\xi)Y(\xi)\,d\xi.$$

Then

$$LY = \int_{-\infty}^{\infty} \lambda \mathcal{Y}(x, \lambda)\, dP(\lambda) G(\lambda),$$

which is equivalent to

$$LY(x) = \int_{-\infty}^{\infty} \lambda\, dE_\lambda\, Y(x).$$

The resolvent operator also has the standard representation. If λ_0 is complex,

$$(L-\lambda_0 I)^{-1} F(x) = \int_{-\infty}^{\infty} \frac{1}{\lambda-\lambda_0}\, dE_\lambda\, F(x).$$

It is apparent that λ_0 is in the spectrum of L if and only if it is in the support of dE_λ or $dP(\lambda)$.

Example. (The Legendre Squared Problem.) The square of the Legendre operator is

$$Ly = [(1-x^2)^2 y'']'' - 2[(1-x^2)y']'.$$

We put this in system format by setting $y_1 = y$, $y_2 = y'$, $y_3 = -((1-x^2)^2 y'')' - 2(1-x^2)y'$, $y_4 = (1-x^2)y''$. Then $(L-\lambda)y = 0$ becomes

$$\begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}' = \left[\lambda \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2(1-x^2) & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/(1-x^2)^2 \end{pmatrix} \right] \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}.$$

With $\lambda = 0$ there are four solutions, all of which are square integrable over $(-1, 1)$. They are

$$Y_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} \ln(1+x) \\ (1+x)^{-1} \\ 0 \\ -(1-x)^{-2} \end{pmatrix}, \quad Y_3 = \begin{pmatrix} \ln(1-x) \\ -(1-x)^{-1} \\ 0 \\ -(1+x)^{-2} \end{pmatrix}$$

and Y_4, whose initial component is

$$y_1 = \int^x (1-t)^{-1} \int^t [\ln(1-s)/(1+s)^2]\, ds\, dt,$$

but nonetheless is square integrable. Each generates boundary conditions at ± 1. We content ourselves with displaying those satisfied by the Legendre polynomials. They are in terms of scalar y,

$$\lim_{x\to -1} [(1-x^2)^2 y''' - 4x(1-x^2)y'' - 2(1-x^2)y'] = 0,$$

$$\lim_{x\to -1} (1+x)^2(-y' + (1-x)y'') = 0,$$

$$\lim_{x\to 1} [-(1-x^2)^2 y''' + 4x(1-x^2)y'' + 2(1-x^2)y'] = 0,$$

$$\lim_{x\to 1} (1-x)^2(y' + (1+x)y'') = 0.$$

For system Y, they are, of course, $\lim Y_j^* J Y$, where $j = 1, 2$ as $x \to 1$, $j = 1, 3$ as $x \to -1$.

REFERENCES

[1] F. V. ATKINSON, *Discrete and Continuous Boundary Value Problems*, Academic Press, New York, 1964.
[2] F. BRAUER, *Spectral theory for linear systems of differential equations*, Pacific J. Math., 10 (1960), pp. 17-34.
[3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, II, Wiley-Interscience, New York, 1963.
[5] W. N. EVERITT, *Integrable-square solutions of ordinary differential equations* II, Quart. J. Math. Oxford Ser. 1, 13 (1962), pp. 217-220.
[6] D. B. HINTON AND J. K. SHAW, *Titchmarsh's λ-dependent boundary conditions for Hamiltonian systems*, in Lecture Notes in Math. 964, Springer-Verlag, Berlin, New York, 1982.
[7] ———, *Hamiltonian systems of limit point or limit circle type both end points singular*, J. Differential Equations, 50 (1983), pp. 444-464.
[8] ———, *On boundary value problems for Hamiltonian systems with two singular points*, SIAM J. Math. Anal., 15 (1984), pp. 272-286.
[9] A. M. KRALL, *Orthogonal polynomials satisfying fourth order differential equations*, Proc. Roy. Soc. Edinburgh, 87 (1981), pp. 271-288.
[10] ———, *Applied Analysis*, D. Reidel, Dordrecht, the Netherlands, 1986.
[11] ———, *$M(\lambda)$ Theory for singular Hamiltonian systems with one singular point*, SIAM J. Math. Anal., this issue (1989), pp. 664-700.

SIAM J. MATH ANAL.
Vol. 20, No. 3, pp. 716-726, May 1989

014

EXISTENCE AND UNIQUENESS THEOREMS FOR THREE-POINT BOUNDARY VALUE PROBLEMS*

A. R. AFTABIZADEH,† CHAITAN P. GUPTA,‡ AND JIAN-MING XU†

Abstract. Existence and uniqueness theorems for third-order boundary value problems are studied. The methods used are the Leray-Schauder continuation theorem and Wirtinger-type inequalities.

Key words. third-order boundary value problems, Leray-Schauder continuation theorem, Wirtinger-type inequalities, sandwich beam

AMS(MOS) subject classifications. 34B10, 34B15

1. Introduction.

The existence and uniqueness of third-order boundary value problems deserve a good deal of attention, since they occur in a wide variety of applications. For example, a three-layer beam is formed by parallel layers of different materials. For an equally loaded beam of this type, Krajcinovic [6] has shown that the deflection ψ is governed by an ordinary third-order linear differential equation

$$\psi''' - K^2\psi' + a = 0,$$

where K^2 and a are physical parameters depending on the elasticity of the layers. The condition of zero moment at the free ends implies the boundary conditions,

$$\psi'(0) = \psi'(1) = 0,$$

and symmetry yields the third boundary condition

$$\psi(1/2) = 0.$$

For recent results concerning third-order boundary value problems, we refer the readers to [1]-[5], [8]-[10].

The purpose of this paper is to study existence and uniqueness results for nonlinear third-order boundary value problems

$$u''' + f(u')u'' = g(x, u, u', u'') + e(x) \tag{1.1}$$

$$u'(0) = u'(1) = u(\eta) = 0, \quad 0 \le \eta \le 1, \tag{1.2}$$

and

$$u''' = g(x, u, u', u'') + e(x) \tag{1.3}$$

$$u'(0) = u''(1) = u(\eta) = 0, \quad 0 \le \eta \le 1, \tag{1.4}$$

* Received by the editors June 15, 1987; accepted for publication June 30, 1988. The first author's research was partially supported by U.S. Army Research Office grant DAAG29-84-G-0034.

†Department of Mathematics, Ohio University, Athens, Ohio 45701.

‡Department of Mathematics, Northern Illinois University, DeKalb, Illinois 60115.

where $f \in C(R, R)$ and $g : [0,1] \times R^3 \to R$ satisfies Carathéodory's conditions, that is

(1) for almost everywhere $x \in [0,1]$, the function $u \in R^3 \to g(x,u) \in R$ is continuous;

(2) for every $u \in R^3$, the function $x \in [0,1] \to g(x,u) \in R$ is measurable;

(3) for every $r > 0$, there is a real valued function $g_r(x) \in L^1[0,1]$ such that for almost everywhere $x \in [0,1]$, $|g(x,u)| \le g_r(x)$ wherever $\|u\| \le r$.

The boundary value problems (1.1), (1.2) and (1.3), (1.4) are in the form of operator equations

$$Lu + Nu = w, \quad \text{where} \quad L : D(L) \subset X \to Y$$

is a linear operator, $N : X \to Y$ is a nonlinear operator and X, Y are suitably Banach spaces in duality (denoted by $(\ ,\ )$). Clearly the linear operator L in (1.1), (1.2) or (1.3), (1.4) is given by

$$Lu = u''',$$

where the boundary conditions (1.2) or (1.3) are used to define the domain, $D(L)$, of L. Our study is motivated by the observation that although it is not possible to obtain necessary a priori estimates to use Leray-Schauder continuation theorem using (Lu, u), it is possible to obtain the necessary a priori estimates using (Lu, u'). In fact, we use (Lu, u') and Wirtinger-type inequalities to obtain the needed a priori estimates to apply Leray-Schauder continuation theorem. Accordingly, we believe that our methods to study the boundary value problems (1.1), (1.2) and (1.3), (1.4) are natural and different than those used in [1]-[5], [8]-[10].

In §§2 and 3 we present some existence and uniqueness results for problems (1.1), (1.2) and (1.3), (1.4), and in §4 we compare our results with the results given in [3], [10]. First we present some results, which help to simplify the proofs of our main results. Let us define

$$\|u\|_\infty = \sup_{0 \le x \le 1} |u(x)|, \quad \text{and} \quad \|u\|_2^2 = \int_0^1 u^2(x)\,dx.$$

LEMMA 1.1. *If* $u(x) \in C^1[0,1]$ *and* $u(0) = 0$, *then* $\|u\|_2^2 \le (4/\pi^2)\|u'\|_2^2$.

LEMMA 1.2. *If* $u(x) \in C^1[0,1]$ *and* $u(0) = u(1) = 0$, *then* $\|u\|_2^2 \le (1/\pi^2)\|u'\|_2^2$.

LEMMA 1.3. *Let* $M_\eta = \max\{\eta, 1-\eta\}$, $0 \le \eta \le 1$. *If* $u(\eta) = 0$, *then*

$$\|u\|_2^2 \le \frac{4}{\pi^2} M_\eta^2 \|u'\|_2^2.$$

Proof. Since $u(\eta) = 0$, then by Lemma 1.1,

$$\int_0^\eta u^2(x)\,dx \le \frac{4}{\pi^2}\eta^2 \int_0^\eta [u'(x)]^2\,dx$$

and

$$\int_\eta^1 u^2(x)\,dx \le \frac{4}{\pi^2}(1-\eta)^2 \int_\eta^1 [u'(x)]^2\,dx.$$

Thus

$$\int_0^1 u^2(x)\,dx \le \frac{4}{\pi^2}\eta^2 \int_0^\eta [u'(x)]^2\,dx + \frac{4}{\pi^2}(1-\eta)^2 \int_\eta^1 [u'(x)]^2\,dx,$$

or

$$\|u\|_2^2 \leq \frac{4}{\pi^2} M_\eta^2 \|u'\|_2^2.$$

LEMMA 1.4. *If* $u(0) = u(1) = 0$, *then*

$$\|u\|_\infty \leq \frac{1}{2} \int_0^1 |u'(x)|\, dx.$$

Let us define the space $H^3(0,1)$ by

$$H^3(0,1) = \left\{ u \in [[0,1], R] \ : \ \frac{d^j u}{dx^j} \ \text{is absolutely continuous on } [0,1] \ \text{for } j = 0,1,2, \text{ and } \frac{d^3 u}{dx^3} \in L^2([0,1]) \right\}.$$

with the usual inner product and the corresponding norm $|\cdot|_{H^3}$. We define a linear operator

$$L \ : \ D(L) \subset C^2[0,1] \to L^1[0,1]$$

by setting

$$D(L) = \{u \in H^3(0,1) \ : \ u \quad \text{satisfies (1.2) or (1.4)}\}$$

and for $u \in D(L)$,

$$Lu = \frac{d^3 u}{dx^3}.$$

LEMMA 1.5. *ker* $L = \{0\}$.

2. Existence results. In this section we apply the results of §1 and the version of Leray-Schauder continuation theorem given by Mawhin in Corollary IV.7 [7] to obtain the existence of a solution for the boundary value problems (1.1), (1.2) and (1.3), (1.4).

THEOREM 2.1. *Let* $g : [0,1] \times R^3 \to R$ *satisfy Carathéodory's conditions, and* $f \in C(R,R)$. *Assume that*

(i) *There exist functions* $a(x) \in C^1[0,1], b(x), c(x) \in C[0,1]$, $d(x) \in L^1[0,1]$, *and real numbers* a_0, b_0, $c_0 \in R$ *such that*

$$a'(x) \leq a_0, \quad b(x) \geq -b_0, \quad c(x) \geq -c_0 \quad \text{for a.e. } x \in [0,1],$$

and for every u, v, $w \in R$, *a.e.* $x \in [0,1]$

$$g(x,u,v,w)v \geq a(x)vw + b(x)v^2 + c(x)|uv| + d(x)|v|;$$

(ii) *There exist* $\alpha \in C\big[[0,1] \times R^2, R\big]$ *and* $\beta \in L^1[0,1]$ *such that*

$$\big|g(x,u,v,w)\big| \leq \big|\alpha(x,u,v)\big||w|^2 + \beta(x)$$

for every $u, v, w \in R$, *and a.e.* $x \in [0,1]$.
Then for every $e(x) \in L^1[0,1]$, *the problem* (1.1) *with* (1.2) *has at least one solution if*

$$(a_0 + 2b_0)\pi + 4M_\eta c_0 < 2\pi^3,$$

where $M_\eta = \max(\eta, 1-\eta)$.

Proof. Let X denote the Banach space $C^2[0,1]$ and Y denote the Banach space $L^1[0,1]$ with its usual norm. Also for $u \in X$, $v \in L^1[0,1]$ let

$$(u,v) = \int_0^1 u(x)v(x)\,dx,$$

denote the duality pairing. We define a linear mapping $L : D(L) \subset X \to Y$ by setting

$$D(L) = \left\{ \begin{array}{l} u \in X \mid u'' \text{ absolutely continuous on } [0,1], \\ \qquad\qquad \text{and } u'(0) = u'(1) = u(\eta) = 0 \end{array} \right\},$$

and for $u \in D(L)$,

$$Lu = u'''.$$

We also define a nonlinear mapping $N : X \to Y$ by setting

$$(Nu)(x) = f(u'(x))u''(x) - g\big(x, u(x), u'(x), u''(x)\big).$$

We note that N is a bounded, continuous mapping. Next, it is easy to see that the linear mapping $L : D(L) \subset X \to Y$, defined above, is a one-to-one mapping. Also the linear mapping $K : Y \to X$, defined, for $y \in Y$, by

$$(Ky)(x) = \int_\eta^x \int_0^t \int_0^s y(\tau)\,d\tau\,ds\,dt + \frac{(\eta^2 - x^2)}{2} \int_0^1 \int_0^t y(\tau)\,d\tau\,dt$$

is such that for $y \in Y$, $Ky \in D(L)$ and $LKy = y$ and for $u \in D(L)$, $KLu = u$. Furthermore, it follows easily by using the Arzela–Ascoli theorem that K maps a bounded subset of Y into relatively compact subsets of X. Hence, $KN : X \to X$ is a compact mapping.

We next note that $u \in C^2[0,1]$ is a solution of the boundary value problem (1.1), (1.2) if and only if u is a solution of the operator equation

$$Lu + Nu = e.$$

Now, the operator equation $Lu + Nu = e$ is equivalent to the equation

$$u + KNu = Ke.$$

We now apply the Leray-Schauder continuation theorem (see, e.g., [7, Cor. IV.7]) to obtain the existence of a solution for $u + KNu = Ke$ or equivalently to the boundary value problem (1.1) and (1.2).

To do this it is suffices to verify that the set of all possible solutions of the family of equations

$$(2.1) \qquad \left\{ \begin{array}{ll} u''' + \lambda f(u')u'' = \lambda g(x,u,u',u'') + \lambda e(x), & x \in (0,1) \\ u(\eta) = u'(0) = u'(1) = 0, & 0 \le \eta \le 1, \end{array} \right.$$

is, a priori, bounded in $C^2[0,1]$ by a constant independent of $\lambda \in [0,1]$.

Let $u(x)$ be a possible solution of (2.1) for some $\lambda \in [0,1]$. Since $u'(0) = u'(1) = 0$, then by Lemma 1.2,

$$\|u'\|_2^2 \leq \frac{1}{\pi^2}\|u''\|_2^2,$$

and from Lemma 1.4

$$\|u'\|_\infty \leq \frac{1}{2}\|u''\|_2.$$

From $u(\eta) = 0$ and Lemma 1.3, we have

$$\|u\|_2^2 \leq \frac{4}{\pi^2}M_\eta^2\|u'\|_2^2 \leq \frac{4}{\pi^4}M_\eta^2\|u''\|_2^2.$$

On multiplying the equation (2.1) by u' and integrating from zero to 1, we have

$$\int_0^1 u'u'''\,dx + \lambda\int_0^1 f(u')u'u''\,dx = \lambda\int_0^1 g(x,u,u',u'')u'\,dx + \lambda\int_0^1 e(x)u'\,dx.$$

Since $u'(0) = u'(1) = 0$, it follows that

$$\int_0^1 f(u')u'u''\,dx = 0,$$

and then from the condition (i), we have

$$-\int_0^1 [u''(x)]^2\,dx \geq \lambda\int_0^1 a(x)u'u''\,dx + \lambda\int_0^1 b(x)[u'(x)]^2\,dx + \lambda\int_0^1 c(x)|uu'|\,dx + \lambda\int_0^1 d(x)|u'|\,dx + \lambda\int_0^1 e(x)u'\,dx$$

$$-\int_0^1 [u''(x)]^2\,dx \geq -\frac{\lambda}{2}\int_0^1 a'(x)[u'(x)]^2\,dx + \lambda\int_0^1 b(x)[u'(x)]^2\,dx + \lambda\int_0^1 c(x)|uu'|\,dx + \lambda\int_0^1 d(x)|u'|\,dx + \lambda\int_0^1 e(x)u'\,dx$$

or

$$\|u''\|_2^2 \leq \lambda\Big[\Big(\frac{a_0}{2} + b_0\Big)\|u'\|_2^2 + c_0\|u\|_2\|u'\|_2 + \|d\|_1\|u'\|_\infty + \|e\|_1\|u'\|_\infty\Big].$$

Hence from the estimates following (2.1), we obtain

$$\|u''\|_2 \leq \frac{\pi^3\big(\|d\|_1 + \|e\|_1\big)}{2\pi^3 - \pi(a_0 + 2b_0) - 4M_\eta c_0} = \rho.$$

Thus

$$\|u'\|_\infty \leq \rho \quad \text{and} \quad \|u\|_\infty \leq \rho.$$

Now, let us assume that

$$M_\rho = \max\big|f(v)\big|, \quad v \in [-\rho, \rho],$$

then

$$|u'''| \leq |f(u')||u''| + |g(x, u, u', u'')| + |e(x)|,$$

and using condition (ii) we have

$$\begin{aligned}\|u'''\|_1 &\leq M_\rho \|u''\|_2 + \int_0^1 |g(x, u, u', u'')|\, dx + \|e\|_1 \\ &\leq M_\rho \rho + \int_0^1 |\alpha(x, u, u')||u''|^2\, dx + \|\beta\|_1 + \|e\|_1 \\ &\leq \rho M_\rho + \rho^2 K_\rho + \|\beta\|_1 + \|e\|_1 \quad = \rho_1,\end{aligned}$$

where

$$K_\rho = \max |\alpha(x, u, v)| \quad \text{on} \quad [0,1] \times [-\rho, \rho] \times [-\rho, \rho].$$

Further, since $u'(0) = u'(1) = 0$, there is an $\xi \in [0,1]$ such that $u''(\xi) = 0$ and $u''(x) = \int_\xi^x u'''(t)\, dt$, for $x \in [0,1]$. It follows that

$$\|u''\|_\infty \leq \|u'''\|_1 \leq \rho_1.$$

All of these considerations imply that there is a constant C, independent of $\lambda \in [0,1]$ such that

$$\|u\|_{C^2[0,1]} \leq C.$$

This completes the proof of this theorem. □

Essentially the same reasoning establishes Theorem 2.2.

THEOREM 2.2. *Suppose all conditions of Theorem* 2.1 *hold true, except in condition* (i) *we assume that* $a(x) \in C[0,1]$ *and* $a(x) \geq -a_0$, *and*

$$v.g(x, u, v, w) \geq a(x)|vw| + b(x)v^2 + c(x)|uv| + d(x)|v|,$$

then problem (1.1), (1.2) *has at least one solution provided*

$$a_0\pi^2 + b_0\pi + 2c_0 M_\eta < \pi^3$$

Remark 2.1. Theorems 2.1 and 2.2 give the solvability of problem (1.1), (1.2) for every given $e(x)$ in $L^1[0,1]$, it is obvious that Theorem 2.1 also gives the solvability of the equation (1.1) with inhomogeneous boundary conditions

$$u'(0) = A_1, \quad u'(1) = A_2, \quad u(\eta) = A_3.$$

COROLLARY 2.1. *Let* $g : [0,1] \times R^3 \to R$ *satisfy Carathéodory's conditions,* $f : R \to R$ *be continuous and assume that for almost everywhere* $x \in [0,1]$, *the function* $g(x, u, v, w)$ *is continuously differentiable with respect to* u, v, *and* w. *Suppose that there exist real numbers* a_0, b_0, *and* c_0 *with* $a_0\pi^2 + b_0\pi + 2c_0 M_\eta < \pi^3$ *such that*

$$\frac{\partial g}{\partial u}(x, u, v, w) \geq -c_0, \quad \frac{\partial g}{\partial v}(x, 0, v, w) \geq -b_0 \quad \left|\frac{\partial g}{\partial w}(x, 0, 0, w)\right| \leq a_0, \tag{2.2}$$

for almost everywhere $x \in [0,1]$ *and all* u, v, $w \in R$. *Suppose further that there exists a continuous function* $\alpha : [0,1] \times R^2 \to R$ *and* $\beta(x) \in L^1[0,1]$ *such that*

$$|g(x,u,v,w)| \leq |\alpha(x,u,v)||w|^2 + \beta(x), \tag{2.3}$$

for every u, v, $w \in R$ *and for almost everywhere* $x \in [0,1]$.

Then for every given $e(x) \in L^1[0,1]$ *the boundary value problem* (1.1), (1.2) *has a solution.*

We can use the same method as in Theorem 2.1 to prove the following theorem for the boundary value problem (1.3), (1.4).

THEOREM 2.3. *Let* $g : [0,1] \times R^3 \to R$ *satisfy Carathéodory's conditions. Assume that*

(i) *There exist functions* $a(x)$, $b(x)$, *and* $c(x) \in C[0,1]$, $d(x) \in L^1[0,1]$, *and real numbers* $a_0, b_0, c_0 \in R$ *such that for every* $x \in [0,1]$

$$a(x) \geq -a_0, \quad b(x) \geq -b_0, \quad c(x) \geq -c_0$$

and for every u, v, $w \in R$, *almost everywhere* $x \in [0,1]$

$$g(x,u,v,w)v \geq a(x)|vw| + b(x)v^2 + c(x)|uv| + d(x)|v|;$$

(ii) *There exist* $\alpha \in C[[0,1] \times R^2, R]$ *and* $\beta \in L^1[0,1]$ *such that*

$$|g(x,u,v,w)| \leq |\alpha(x,u,v)||w|^2 + \beta(x)$$

for every u, v, $w \in R$, *almost everywhere* $x \in [0,1]$.
Then for every $e(x) \in L^1[0,1]$, *problem* (1.3)-(1.4) *has at least one solution if*

$$2\pi^2 a_0 + 4\pi b_0 + 8M_\eta c_0 < \pi^3,$$

where $M_\eta = \max(\eta, 1-\eta)$.

COROLLARY 2.2. *Suppose that all conditions of Corollary* 2.1 *hold true, except that the condition* $a_0\pi^2 + b_0\pi + 2c_0M_\eta < \pi^3$ *is replaced by*

$$2\pi^2 a_0 + 4\pi b_0 + 8M_\eta c_0 < \pi^3,$$

then boundary value problem (1.3), (1.4) *has a solution.*

3. Uniqueness results. In this section we discuss existence of a unique solution for the boundary value problems

$$u''' + Au'' = g(x,u,u',u'') + e(x) \tag{3.1}$$

$$u(\eta) = u'(0) = u'(1) = 0, \tag{3.2}$$

and

$$u''' = g(x,u,u',u'') + e(x) \tag{3.3}$$

$$u(\eta) = u'(0) = u''(1) = 0, \tag{3.4}$$

where A is a constant and $g(x, u, v.w)$ satisfies Carathéodory's conditions, and $e(x) \in L^1[0,1]$.

THEOREM 3.1. *Let $g : [0,1] \times R^3 \to R$ satisfy Carathéodory's conditions, and A be a constant. Assume there exist functions $a(x) \in C^1[0,1]$, $b(x)$, $c(x) \in C[0,1]$, and constants a_0, b_0, $c_0 \in R$ such that for almost everywhere $x \in [0,1]$*

$$a'(x) \le a_0,\ b(x) \ge -b_0,\ c(x) \ge -c_0$$

and for every u_i, v_i, $w_i \in R$, $i = 1, 2$, and almost everywhere $x \in [0,1]$

$$\begin{aligned}\big(g(x,u_1,v_1,w_1) - g(x,u_2,v_2,w_2)\big)(v_1 - v_2) \ge & a(x)(w_1 - w_2)(v_1 - v_2) \\ & + b(x)(v_1 - v_2)^2 + c(x)|u_1 - u_2||v_1 - v_2|.\end{aligned}$$

Then for every $e(x) \in L^1[0,1]$, problem (3.1), (3.2) has a unique solution provided

$$(a_0 + 2b_0)\pi + 4c_0 M_\eta < 2\pi^3.$$

Proof. Let us assume that u_1 and u_2 are two solutions of (3.1)-(3.2), then

$$(u_1 - u_2)''' + A(u_1 - u_2)'' = g(x, u_1, u_1', u_1'') - g(x, u_2, u_2', u_2'') \tag{3.5}$$

and

$$(u_1 - u_2)(\eta) = 0,\ (u_1 - u_2)'(0) = 0,\ (u_1 - u_2)'(1) = 0. \tag{3.6}$$

On multiplying (3.5) by $(u_1 - u_2)'$ and integrating for 0 to 1, we have

$$-\int_0^1 \big[(u_1 - u_2)''\big]^2\,dx = \int_0^1 \big[g(x, u_1, u_1', u_1'') - g(x, u_2, u_2', u_2'')\big](u_1 - u_2)'\,dx.$$

Let $y = u_1 - u_2$, then from the condition (i)

$$\begin{aligned}-\int_0^1 [y'']^2\,dx &\ge \int_0^1 a(x)y''y'\,dx + \int_0^1 b(x)[y']^2\,dx + \int_0^1 c(x)|y||y'|\,dx \\ &\ge -\frac{1}{2}\int_0^1 a'(x)(y')^2\,dx + \int_0^1 b(x)(y')^2\,dx + \int_0^1 c(x)|y||y'|\,dx\end{aligned}$$

or

$$\begin{aligned}\int_0^1 (y'')^2\,dx &\le \Big(\frac{1}{2}a_0 + b_0\Big)\int_0^1 (y')2\,dx + c_0\Big[\int_0^1 y^2\,dx\Big]^{1/2}\Big[\int_0^1 (y')^2\,dx\Big]^{1/2} \\ &\le \frac{a_0/2 + b_0}{\pi^2}\int_0^1 (y'')^2\,dx + \frac{2M_\eta c_0}{\pi^3}\int_0^1 (y'')^2\,dx\end{aligned}$$

or $\|y''\|_2{}^2 \le 0$. From Lemma 1.2, it follows that $\|y'\|_2^2 \le 0$. Since

$$\|y\|_\infty \le \|y'\|_2 \le 0,$$

then $y(x) = 0$, and hence $u_1(x) = u_2(x)$ for almost everywhere $x \in [0,1]$. But $H^3(0,1) \subset C^2[0,1]$, which implies $u_1(x) = u_2(x)$ for every $x \in [0,1]$. The proof is complete. □

THEOREM 3.2. *Let $g : [0,1] \times R \to R$ satisfy Carathéodory's conditions, and A be a constant. Assume there exist functions $a(x)$, $b(x)$, $c(x) \in C[0,1]$, and constants a_0, b_0, $c_0, \in R$ such that for almost everywhere $x \in [0,1]$*

$$a(x) \geq -a_0, \quad b(x) \geq -b_0, \quad c(x) \geq -c_0$$

and for every u_i, v_i, $w_i \in R$, $i = 1,2$ and almost everywhere $x \in [0,1]$

$$\Big(g(x,u_1,v_1,w_1) - g(x,u_2,v_2,w_2)\Big)(v_1 - v_2) \geq a(x)|w_1 - w_2||v_1 - v_2| + b(x)(v_1 - v_2)^2 + c(x)|u_1 - u_2|\,|v_1 - v_2|.$$

Then for every $e(x) \in L^1[0,1]$, the problem (3.1), (3.2) *has a unique solution provided*

$$a_0\pi^2 + b_0\pi + 2c_0M_\eta < \pi^3.$$

THEOREM 3.3. *Let $g : [0,1] \times R^3 \to R$ satisfies Carathéodory's conditions. Assume there exist functions $a(x)$, $b(x)$, $c(x) \in C[0,1]$ and constants a_0, b_0, $c_0 \in R$ such that for almost everywhere $x \in [0,1]$*

$$a(x) \geq -a_0, \quad b(x) \geq -b_0, \quad |c(x)| \leq c_0$$

and for every u_i, v_i, $w_i \in R$, i=1,2, and almost everywhere $x \in [0,1]$,

$$\big(g(x,u_1,v_1,w_1) - g(x,u_2,v_2,w_2)\big)(v_1 - v_2) \geq a(x)|w_1 - w_2|v_1 - v_2| + b(x)(v_1 - v_2)^2 + c(x)|u_1 - u_2||v_1 - v_2|.$$

Then for every $e(x) \in L^1[0,1]$ the problem (3.3), (3.4) *has a unique solution if*

$$2\pi^2a_0 + 4\pi b_0 + 8M_\eta c_0 < \pi^3.$$

Remark 3.1. We remark that Theorems 3.1–3.3 give uniqueness results for the boundary value problems (3.1), (3.2) and (3.3), (3.4). To obtain existence and uniqueness results for (3.1), (3.2) and (3.3), (3.4) we only need to combine the theorems of §§2 and 3.

4. Examples and comparisons. From Theorem 2.2 it is easy to prove the following corollary.

COROLLARY 4.1. *Suppose all conditions of Theorem* 2.2 *hold true except condition* (ii) *which we replace by*

(ii)′ *There exist $\alpha \in C\big[[0,1] \times R^2, R\big]$, $\gamma \in C\big[[0,1] \times R^3, R\big]$, and $\beta \in L^1[0,1]$ such that*

$$\big|g(x,u,v,w)\big| \leq \alpha(x,u,v)|w|^2 + \gamma(x,u,v,w) + \beta(x)$$

for every u, v, $w \in R$ *and almost everywhere* $x \in [0,1]$, *where* $\gamma(x,u,v,w)$ *is bounded when* (x,u,v) *varies in a bounded set in* $[0,1] \times R^2$. *Then the problem* (1.1), (1.2) *has a solution provided*

$$a_0\pi^2 + b_0\pi + 2c_0M_\eta < \pi^3 .$$

Aftabizadeh and Wiener [1] studied problem (1.1), (1.2) when $f \equiv 0$ and $\eta = 0$. Granas, Guenther, and Lee [3] discussed the existence of solutions of boundary value problem (1.3) ((1.2), or (1.4)) where $g = \varphi + \psi$, they proved the following [3].

The boundary value problem

$$u''' = \varphi(x,u,u',u'') + \psi(x,u,u',u''), \quad u(0) = u'(0) = u'(1) = 0$$

has at least one solution provided the functions φ and ψ are continuous and

(a) $\varphi \cdot v \geq 0$ on $[0,1] \times R^3$

(b) $|\psi| \leq B\big(1 + |u|^\alpha + |v|^\beta + |w|^\gamma\big)$, where $B < \infty$, $0 \leq \alpha, \beta, \gamma < 1$.

(c) $\varphi(x,u,v,w)$ is bounded when (x,u,v) in a bounded set $[0,1] \times R^2$.

Corollary 4.1 covers this results. Indeed it is clear that (b) implies there exists $B' < \infty$, such that for every u, v, $w \in R$, $x \in [0,1]$

$$|\psi(x,u,v,w)| \leq |u| + |v| + |w| + B. \tag{4.1}$$

Then from (a) we have

$$(\varphi + \psi).v \geq \psi.v \geq -|\psi|.|v| \geq -|uv| - v^2 - |vw| - B'|v|. \tag{4.2}$$

Also from (4.1) we have

$$|\varphi + \psi| \leq |\varphi| + |\psi| \leq |w|^2 + \big[|\varphi| + |u| + |v| + B' + 1\big].$$

Hence, if we take $f \equiv 0$, $e \equiv 0$ and $g = \varphi + \psi$ in Corollary 4.1 with $\eta = 0$, then the result of [3] follows.

O'Regan [10] proved the following theorem.

THEOREM O. *Let* $g : [0,1] \times R^3 \to R$ *be continuous.*

(a) *Suppose there is a constant* $M \geq 0$ *such that*

$$pg(x,u,p,0) \geq 0 \quad \textit{for} \quad |p| > M \quad \textit{and} \quad (x,u) \in [0,1] \times R.$$

(b) *Suppose that*

$$|g(x,u,p,q)| \leq A(x,u,p)q^2 + B(x,u,p)$$

where $A(x,u,p), B(x,u,p) \geq 0$ *are functions bounded on bounded* (x,u,p) *sets. Then the boundary value problem*

$$u''' = g(x,u,u',u''), \quad u(\eta) = u'(0) = u'(1) = 0, \quad x \in [0,1],$$

has at least one solution in $C^3[0,1]$.

In general this theorem covers more class of differential equations than our Theorem 2.1 if $f \equiv 0$. Condition (a) is weaker than condition (i) of Theorem 2.1. If

the function $g(x,u,v,w)$ is independent of w, then our results are stronger, as the following example shows.

Example. (Sandwich beam). Consider the differential equation

$$\psi''' = k^2(x,\psi)\psi' - a(x,\psi), \qquad x \in [0,1],$$

with the boundary conditions

$$\psi'(0) = \psi'(1) = \psi(1/2) = 0.$$

O'Regan [10] assumed that $k^2(x,\psi)$ and $a(x,\psi)$ are continuous functions on $[0,1]\times R$. In addition, suppose there exists a constant $L < \infty$ such that

$$\left|\frac{a(x,\psi)}{k^2(x,\psi)}\right| \le L \qquad \text{for } (x,\psi) \in [0,1]\times R.$$

Then $\psi' g(x,\psi,\psi) = \psi'(k^2\psi' - a) > 0$ for $|\psi'| > L$ and $(x,\psi) \in [0,1]\times R$, and so Theorem O implies that the boundary value problem has at least one solution in $C^3[0,1]$.

Now we make the following assumptions on k and a.

Suppose $k^2(x,\psi) \equiv 1$ and $a(x,\psi) = a(x)\psi + b(x)$ then Theorem O does not guarantee this problem has a solution while Theorem 2.1 implies the boundary value problem has a solution. More generally, if we suppose that k, $a \in C\big[[0,1]\times R, R\big]$ and there exists functions $c(x) \in C[0,1]$, $d(x) \in L^1[0,1]$ such that $c(x) > -\pi^3$ and

$$\psi' \,.\, a(x,\psi) \le c(x)|\psi.\psi'| + d(x)|\psi'|,$$

then Theorem 2.1 guarantees this boundary value problem has at least one solution in $C^3[0,1]$.

REFERENCES

[1] A.R. AFTABIZADEH AND J. WIENER, *Existence and uniqueness theorems for third order boundary value problems*, Rend. Sem. Mat. Univ. Padova, 75 (1986), pp. 130-141.

[2] R.P. AGARWAL, *Existence-uniqueness and iterative methods for third order boundary value problems*, J. Comp. Appl. Math., to appear.

[3] A. GRANAS, R. GUENTHER, AND J. LEE, *Nonlinear boundary value problems for ordinary differential equations*, Polish Academy of Science, Poland, 1985.

[4] M. GREGUS, *Third order linear differential equations*, D. Reidel, Dordrecht, the Netherlands, 1987.

[5] J. HENDERSON, *Best interval lengths for boundary value problems for third order Lipschitz equations*, SIAM J. Math. Anal., 18 (1987), pp. 293-305.

[6] D. KRAJCINOVIC, *Sandwich beam analysis*, J. Appl. Mech., 39 (1972), pp. 773-778.

[7] J. MAWHIN, *Topological degree methods in nonlinear boundary value problems*, NSF-CBMS Regional Conference Series in Math. 40, American Mathematical Society, Providence, RI, 1979.

[8] K.N. MURTY AND B.D.C.N. PRASAD, *Three-point boundary value problems existence and uniqueness*, Yokohama Math. J., 29 (1981), pp. 101-105.

[9] ————, *Application of Liapunov theory to three point boundary value problems*, J. Math. Phy. Sci., 19 (1985), pp. 225-234.

[10] D.J. O'REGAN, *Topological transversality: Applications to third order boundary value problems*, SIAM J. Math. Anal., 18 (1987), pp. 630-641.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 727-737, May 1989

015

QUADRATIC BIRTH AND DEATH PROCESSES AND ASSOCIATED CONTINUOUS DUAL HAHN POLYNOMIALS*

MOURAD E. H. ISMAIL†, JEAN LETESSIER,‡ AND GALLIANO VALENT‡

Abstract. Birth and death process polynomials with symmetric quadratic rates are studied. They provide two generalizations of the continuous dual Hahn polynomials. Generating functions and explicit representations are derived. The asymptotic behavior and weight functions of the polynomials under consideration are also determined.

Key words. continuous dual Hahn polynomials, birth and death processes, generating functions

AMS(MOS) subject classifications. 33A65, 42C05, 60K25

1. Introduction. A birth and death process with birth rates $\{\lambda_n\}$ and death rates $\{\mu_n\}$ gives rise to a set of polynomials $\{p_n(x)\}$ defined recursively by

$$p_0(x)=1,\ p_1(x)=(\lambda_0+\mu_0-x)/\mu_1, \tag{1.1}$$

$$-xp_n(x)=\mu_{n+1}p_{n+1}(x)+\lambda_{n-1}p_{n-1}(x)-(\lambda_n+\mu_n)p_n(x). \tag{1.2}$$

It is assumed that

$$\lambda_n>0,\quad \mu_{n+1}>0,\quad n\geqq 0,\quad \mu_0\geqq 0. \tag{1.3}$$

The p_n's are orthogonal with respect to a probability measure $d\psi$ with finite moments. The orthogonality relation is

$$\int_0^\infty p_m(x)\,p_n(x)\,d\psi(x)=\pi_n\delta_{m,n}, \tag{1.4}$$

where

$$\pi_0=1,\quad \pi_n=\lambda_0\lambda_1,\cdots,\lambda_{n-1}/\mu_1\mu_2,\cdots,\mu_n,\quad n>0. \tag{1.5}$$

Karlin and McGregor [8], [9] proved that the transition probability $p_{mn}(t)$, the probability that the system moves from state m to state n in time t, is given by

$$\pi_n p_{mn}(t)=\int_0^\infty e^{-xt}p_m(x)\,p_n(x)\,d\psi(x).$$

Such a measure $d\psi$ is called a spectral measure of the birth and death process.

Birth and death processes with polynomial rates $\{\lambda_n\}$, $\{\mu_n\}$ arise in many fields [5], but there seem to be very few cases known where the polynomials and the spectral measures are known explicitly. A complete analysis of the cases when both λ_n and μ_{n+1}, $n\geqq 0$ are linear in n was completed only recently [4], [7], [20]. The polynomials are associated Laguerre and associated Meixner polynomials.

We will study symmetric birth and death process polynomials with quadratic rates

$$\lambda_n=(n+a)(n+b),\quad n\geqq 0,\quad \mu_n=(n+\alpha)(n+\beta),\quad n>0,\quad \mu_0=0 \text{ or } \mu_0=\alpha\beta. \tag{1.6}$$

* Received by the editors March 10, 1987; accepted for publication (in revised form) June 30, 1988.

† Department of Mathematics, University of South Florida, Tampa, Florida, 33520. This author's research was partially supported by a grant from the National Science Foundation.

‡ Laboratoire de Physique Theorique et Hautes Energies (an affiliate of Centre National de la Recherche Scientifique), Université Paris VII, 75251 Paris Cedex 05, France.

When α or $\beta = 0$ the polynomials reduce to the dual Hahn polynomials in Askey's tableaux; see [1], [3], and [10]. This case has also been treated in [11]. The cases $ab = 0$ are in [12]. We will call these polynomials the associated continuous dual Hahn polynomials. In § 2 we find generating functions for the polynomials and for the numerator polynomials of the corresponding J fraction. We also apply Darboux's asymptotic method to these generating functions and determine the main term in the asymptotic expansion of the polynomials under consideration. In § 3 we find the continued fraction whose denominators are the associated continuous dual Hahn polynomials. The spectral measure in this case turns out to be unique, so the continued fraction is the Stieltjes transform of the spectral measure. The Stieltjes transform is then inverted and the spectral measure is found. In § 4 we obtain explicit representations for our polynomials and use the symmetry of the polynomials in the parameters α, β to obtain transformation formulas involving double sums.

We follow the notation, terminology, and methodology in [2] and [14]. In particular, if a sequence of orthogonal polynomials $\{r_n(x)\}$ satisfies a three-term recurrence relation

$$r_{n+1}(x) = (a_n x + b_n) r_n(x) - d_n r_{n-1}(x),$$

then the associated polynomials $\{r_n(x; c)\}$ are given initially by

$$r_0(x; c) = 1, \; r_1(x; c) = a_c x + b_c$$

and are defined recursively by

$$r_{n+1}(x; c) = (a_{n+c} x + b_{n+c}) r_n(x; c) - d_{n+c} r_n(x; c), \qquad n > 0,$$

provided that the coefficients a_{n+c}, b_{n+c}, d_{n+c} are well defined and $r_n(x; c)$ has precise degree n. The interesting cases arise when c is not a positive integer. For references on associated polynomials, see [2] and [4]. Recently Wimp [19] made a detailed study of the associated Jacobi polynomials, applying the approach used in [4].

We will follow the standard hypergeometric function notation as in [6], [15], and [17] and the terminology of asymptotic analysis as in Olver [13].

2. Generating functions. We will treat two distinct cases according to whether or not μ_0 vanishes. For convenience we introduce an auxiliary parameter η in the following way.

$$\text{Case I: } \eta = 0 \text{ and } \mu_0 = 0 \qquad \text{Case II: } \eta = 1 \text{ and } \mu_0 = \alpha\beta.$$

Let $\{P_n(x)\}$ (or $\{P_n(x; a, b, \alpha, \beta, \eta)\}$) be the corresponding orthogonal polynomials and set

$$G(x, w) = \sum_0^\infty P_n(x) w^n. \tag{2.1}$$

Multiplying the recurrence relation (1.2) by w^{n+1} and applying the initial conditions (1.1), we derive the differential equation

$$\begin{aligned} w(1-w)^2 G'' + (1-w)[1+\alpha+\beta-(a+b+1)w]G' \\ + [x + (w-1)(ab - \alpha\beta/w)]G = \alpha\beta(1-w)^\eta / w, \end{aligned} \tag{2.2}$$

where $'$ denotes differentiation with respect to w. We then observe that the singularities of the above differential equation are $w = 0, 1, \infty$ and that all three are regular singular points. This means that if H is a solution of the homogeneous equation corresponding

to (2.2) then we can find parameters μ and ν such that the function $F(w) = w^{-\mu}(1-w)^{-\nu}H(w)$ satisfies the hypergeometric differential equation

$$w(1-w)F'' + [C-(1+A+B)w]F' - ABF = 0. \tag{2.3}$$

We make the choices

$$\mu = -\beta, \qquad \nu = \gamma - (\gamma^2 - x)^{1/2} \quad \text{where } \gamma = (1+\alpha+\beta-a-b)/2, \tag{2.4}$$

and find

$$A = -(\gamma^2-x)^{1/2} + (1+\alpha-\beta+b-a)/2, \qquad B = A+a-b, \qquad C = 1+\alpha-\beta. \tag{2.5}$$

At this stage the square root appearing in ν in (2.4) has no particular branch associated with it. Later we will choose the branch that makes $\operatorname{Re}(\gamma^2-x)^{1/2} > 0$. Two linearly independent solutions of the homogeneous differential equation corresponding to (2.2) are given by the functions G_1 and G_2 defined below.

$$G_1^{\alpha,\beta}(w) = w^{-\beta}(1-w)^{\gamma-(\gamma^2-x)^{1/2}} \qquad {}_2F_1(A, B; C; w)$$

$$G_2^{\alpha,\beta}(w) = G_1^{\beta,\alpha}(w).$$

The Wronskian of G_1 and G_2 is $C_1 w^{-\alpha-\beta-1}(1-w)^{\alpha+\beta-a-b}$, C_1 being a constant. Therefore the solution of (2.2), which is analytic at $w=0$ and $G(x,0)=1$, is given by

$$\begin{aligned} G(x, w) = \frac{\alpha\beta}{\alpha-\beta}\int_0^w u^{\alpha+\beta-1}(1-u)^{\eta-2\gamma-1}[G_1^{\alpha,\beta}(w)G_2^{\alpha,\beta}(u) \\ - G_1^{\alpha,\beta}(u)G_2^{\alpha,\beta}(w)]\,du. \end{aligned} \tag{2.6}$$

We next determine the main term in the asymptotic development of $P_n(x)$ using the asymptotic method of Darboux which states that if

$$f(z) = \sum_0^\infty f_n z^n \text{ and } g(z) = \sum_0^\infty g_n z^n$$

are analytic in $|z| < r$, and $f(z) - g(z)$ is continuous on $|z| = r$ then $f_n = g_n + o(r^{-n})$ as $n \to \infty$. Thus the asymptotics of the coefficients in the Taylor series expansion of a function $f(z)$ analytic in a neighborhood of $z=0$ are determined by the main term in the singular part of $f(z)$ at the closest singularities to the origin. For details see Olver [13] and Szegö [18]. It is clear that the ${}_2F_1$'s appearing in (2.6) are defined when their argument equals unity if $\operatorname{Re}\{(\gamma^2-x)^{1/2}\} > 0$. Therefore in this case

$$(1-w)^{-\gamma+\sqrt{\gamma^2-x}}G(x, w)$$

is an analytic function of w in $|w| \leqq 1$. Let $\mathscr{A}$ denote the limit of the above function as $w \to 1$. The smallest exponent of $1-w$ in the expansion of $G(x, w)$ around $w=1$ is $\gamma - (\gamma^2-x)^{1/2}$. We now apply Darboux's asymptotic method. Thus the dominant term in the asymptotic expansion of $P_n(x)$ equals the coefficient of w^n in $\mathscr{A}(1-w)^{\gamma-\sqrt{\gamma^2-x}}$. We then use the binomial theorem to find the latter coefficient to be

$$\mathscr{A}\frac{\Gamma(n-\gamma+\sqrt{\gamma^2-x})}{\Gamma(\sqrt{\gamma^2-x}-\gamma)\Gamma(n+1)}$$

which is asymptotic to $n^{-\gamma-1+\sqrt{\gamma^2-x}}/\Gamma(-\gamma+\sqrt{\gamma^2-x})$. This establishes the asymptotic

formula

$$P_n(x)\approx\frac{\alpha\beta n^{-1-\gamma-(\gamma^2-x)^{1/2}}}{(\beta-\alpha)\Gamma\{(\gamma^2-x)^{1/2}-\gamma\}}\int_0^1(1-u)^{\eta-\gamma-1-\sqrt{\gamma^2-x}}$$
$$\cdot\left[u^{\alpha-1}\frac{\Gamma(1+\beta-\alpha)\Gamma(2(\gamma^2-x)^{1/2})}{\Gamma(1-A)\Gamma(1-B)}{}_2F_1\left(\begin{matrix}A+\beta-\alpha, B+\beta-\alpha\\ 1+\alpha-\beta\end{matrix}\middle| u\right)\right.$$
$$\left.-\text{a similar term with } \alpha \text{ and } \beta \text{ interchanged}\right]du,$$

when $x\notin R$. The integrand in the above asymptotic formula can be greatly simplified by using the connection relation [6, § 2.9, (33)]

$$(1-z)^{h-f-g}{}_2F_1\left(\begin{matrix}h-f, h-g\\ 1+h-f-g\end{matrix}\middle| 1-z\right)$$
$$=\frac{\Gamma(1+h-f-g)\Gamma(1-h)}{\Gamma(1-f)\Gamma(1-g)}{}_2F_1\left(\begin{matrix}f, g\\ h\end{matrix}\middle| z\right)$$
$$+\frac{\Gamma(1+h-f-g)\Gamma(h-1)}{\Gamma(h-f)\Gamma(h-g)}z^{1-h}{}_2F_1\left(\begin{matrix}1+f-h, 1+g-h\\ 2-h\end{matrix}\middle| z\right).$$

The result is

$$P_n(x)\approx\frac{\alpha\beta n^{-1-\gamma-(\gamma^2-x)^{1/2}}}{2(\gamma^2-x)^{1/2}\Gamma\{(\gamma^2-x)^{1/2}-\gamma\}}\int_0^1 u^{\beta-1}(1-u)^{\eta-1-\gamma+\sqrt{\gamma^2-x}}$$
$$\cdot{}_2F_1\left(\begin{matrix}1-A, 1-B\\ 1+2(\gamma^2-x)^{1/2}\end{matrix}\middle| 1-u\right)du,$$

valid for nonreal x. We now apply the Euler type integral representation [15, § 49]

$$(2.7)\quad {}_3F_2\left(\begin{matrix}a_1, a_2, a_3\\ b_1, b_2\end{matrix}\middle| z\right)=\frac{\Gamma(b_2)}{\Gamma(b_2-a_3)\Gamma(a_3)}\int_0^1 t^{a_3-1}(1-t)^{b_2-a_3-1}{}_2F_1\left(\begin{matrix}a_1, a_2\\ b_1\end{matrix}\middle| zt\right)dt$$

and obtain the asymptotic relationship

$$(2.8)\quad P_n(x)\approx\frac{\alpha\beta n^{\sqrt{\gamma^2-x}-1-\gamma}\Gamma(\beta)\Gamma(\eta-\gamma+\sqrt{\gamma^2-x})}{2\sqrt{\gamma^2-x}\Gamma(\sqrt{\gamma^2-x}-\gamma)\Gamma[\eta+\sqrt{\gamma^2-x}+(a+b+\beta-\alpha-1)/2]}$$
$$\cdot{}_3F_2\left(\begin{matrix}a_1, a_2, a_3\\ b_1, b_2\end{matrix}\middle| 1\right)$$

where

$$a_1=(1+a-b+\beta-\alpha)/2+\sqrt{\gamma^2-x},$$
$$a_2=\sqrt{\gamma^2-x}+(1+b-a+\beta-\alpha)/2,$$
$$a_3=\sqrt{\gamma^2-x}+\eta+(a+b-\alpha-\beta-1)/2,$$
$$b_1=1+2\sqrt{\gamma^2-x},$$
$$b_2=\sqrt{\gamma^2-x}+\eta+(a+b+\beta-\alpha-1)/2,$$

as $n\to\infty$ and fixed x with $\operatorname{Im}\{x\}\neq 0$.

3. The spectral measures. The orthonormal polynomials $\{\omega_n(x)\}$ are given by

$$\omega_n(x)=[\mu_1\mu_2,\cdots,\mu_n/\lambda_0\lambda_1,\cdots,\lambda_{n-1}]^{1/2}P_n(x).$$

Therefore

$$\omega_n(x)\approx[\Gamma(a)\Gamma(b)/\Gamma(\alpha+1)\Gamma(\beta+1)]n^{\alpha+\beta+2-a-b}P_n(x),\quad \text{as } n\to\infty.$$

It easily follows from (2.8) and the above asymptotic relationship that the series

$$\sum_0^\infty|\omega_n(x)|^2$$

diverges for some complex x. Now Theorem 2.9 [16, p. 50] implies that the corresponding moment problem is determined, i.e., the spectral measure is unique.

The numerators $\{q_n(x)\}$ of the continued fraction whose partial denominators are the p_n's (of (1.1) and (1.2)) satisfy the recursion (1.2) and the initial conditions

$$q_0(x)=0,\qquad q_1(x)=-1/\mu_1. \tag{3.1}$$

Clearly, $q_n(x)$ is a polynomial of degree $n-1$. Let σ denote the support of the spectral measure $d\psi$. The determinacy of the moment problem ensures the uniform convergence of $q_n(x)/p_n(x)$ to $\int_\sigma(x-t)^{-1}\,d\psi(x)$ on compact subsets of the complex x-plane cut along σ [16, Thm. 2.9, p. 50]. At this stage we need to exhibit the dependence of the polynomials on the parameters involved, so we will use $P_n(x;a,b,\alpha,\beta,\eta)$ instead of just $P_n(x)$ to denote our polynomials and will use a similar notation for the q_n's. We can easily see that

$$\begin{aligned}q_n(x;a,b,\alpha,\beta,\eta)&=[-(\alpha+1)(\beta+1)]^{-1}\\&\quad\cdot P_{n-1}(x;a+1,b+1,\alpha+1,\beta+1,0),\qquad \eta=0,1,\end{aligned} \tag{3.2}$$

follows from (1.1), (1.2), and (3.1). Set

$$X_\eta(x)=\lim_{n\to\infty}q_n(x;a,b,\alpha,\beta,\eta)/P_n(x;a,b,\alpha,\beta,\eta). \tag{3.3}$$

We are now in a position to prove the following theorem.

THEOREM 1. *The functions $X_\eta(x)$ are given by*

$$X_\eta(x)=\frac{\alpha^{-1}{}_3F_2\left(\begin{matrix}1-A,\,1-B,\,\sqrt{\gamma^2-x}-\gamma\\1+2\sqrt{\gamma^2-x},\,b-A+1\end{matrix}\,\middle|\,1\right)}{[\gamma-\sqrt{\gamma^2-x}+\beta(\eta-1)]_3F_2\left(\begin{matrix}1-A,\,1-B,\,\sqrt{\gamma^2-x}-\gamma+\eta\\1+2\sqrt{\gamma^2-x},\,b-A+\eta\end{matrix}\,\middle|\,1\right)},\quad \eta=0,1. \tag{3.4}$$

Proof. The theorem follows from (3.2), (3.3), and (2.8). □

The next step is to invert the Stieltjes transforms $X_\eta(x)$ and find the measures $d\psi$ explicitly via the following inversion formula, which holds when the support of $d\psi$ is contained in a half line

$$\begin{aligned}&\int_{-\infty}^{\infty}\frac{d\psi(t)}{x-t}=F(x),\qquad x\notin\operatorname{supp}\{d\psi\}\\&\text{iff }\psi(t_2)-\psi(t_1)=\lim_{\varepsilon\to0^+}\int_{t_1}^{t_2}\frac{F(t-i\varepsilon)-F(t+i\varepsilon)}{2\pi i}\,dt.\end{aligned} \tag{3.5}$$

The inversion formula (3.5) is usually referred to as the Perron-Stieltjes inversion formula. We now state a technical lemma needed in the inversion of the Stieltjes transforms of the measures $d\psi(x)$.

LEMMA 2. *Assume that* $\operatorname{Re}\{d-a-b\}>0$ *and* $\operatorname{Re}\{a+b+c-d-e\}>0$. *Then*

$$\lim_{z\to 1^-}(1-z)^{a+b+c-d-e}\,{}_3F_2\left(\begin{matrix}a, b, c\\ d, e\end{matrix}\,\middle|\, z\right)=\frac{\Gamma(d)\Gamma(e)\Gamma(a+b+c-d-e)}{\Gamma(a)\Gamma(b)\Gamma(c)} \tag{3.6}$$

Proof. Applying the Pfaff-Kummer transformation [15]

$${}_2F_1(a, b; c; z)=(1-z)^{-a}\,{}_2F_1(a, c-b; c; z/(z-1)) \tag{3.7}$$

to the ${}_2F_1$ in (2.7), then replacing t by $1-t$ we get

$$\begin{aligned}{}_3F_2\left(\begin{matrix}a, b, c\\ d, e\end{matrix}\,\middle|\, z\right)&=\frac{\Gamma(e)(1-z)^{d-a-b}}{\Gamma(c)\Gamma(e-c)}\int_0^1 t^{e-c-1}(1-t)^{c-1}\\ &\quad\cdot\left\{1-\frac{zt}{z-1}\right\}^{d-a-b}{}_2F_1\left(\begin{matrix}d-a, d-b\\ d\end{matrix}\,\middle|\, z(1-t)\right)dt\\ &=\frac{\Gamma(e)(1-z)^{d-a-b}}{\Gamma(c)\Gamma(e-c)}\sum_0^\infty\frac{(d-a)_n(d-b)_n}{n!(d)_n}z^n\int_0^1 t^{e-c-1}(1-t)^{n+c-1}\\ &\quad\cdot\left(1-\frac{zt}{z-1}\right)^{d-a-b}dt\\ &=\frac{\Gamma(e)(1-z)^{d-a-b}}{\Gamma(c)\Gamma(e-c)}\sum_0^\infty\frac{(d-a)_n(d-b)_n}{n!(d)_n}z^n\,{}_2F_1\left(\begin{matrix}e-c, a+b-d\\ e+n\end{matrix}\,\middle|\,\frac{z}{z-1}\right).\end{aligned}$$

In the last step we also used the Euler integral representation for a ${}_2F_1$, i.e., (2.7) with a_1 equal to b_1. Next apply the transformation (3.7) to the ${}_2F_1$ in the last sum to obtain

$$\begin{aligned}(1-z)^{a+b+c-d-e}\,{}_3F_2\left(\begin{matrix}a, b, c\\ d, e\end{matrix}\,\middle|\, z\right)&=\frac{\Gamma(e)}{\Gamma(c)}\sum_0^\infty\frac{(d-a)_n(d-b)_n\Gamma(n+c)}{n!(d)_n\Gamma(e+n)}\\ &\quad\cdot{}_2F_1\left(\begin{matrix}e-c, n+d+e-a-b\\ e+n\end{matrix}\,\middle|\, z\right).\end{aligned} \tag{3.8}$$

We then let $z\to 1^-$ in (3.8) and use Gauss's theorem [15]

$${}_2F_1(\alpha, \beta; \gamma; 1)=\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)/\{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)\},\quad \operatorname{Re}(\gamma-\alpha-\beta)>0.$$

It is easy to justify interchanging the limiting and summation processes. The result is that the right-hand side of (3.8) reduces to

$$\frac{\Gamma(e)}{\Gamma(c)}\sum_0^\infty\frac{(d-a)_n(d-b)_n\Gamma(a+b+c-d-e)}{n!(d)_n\Gamma(a+b-d)}$$

which can be summed by Gauss's theorem. This completes the proof of the lemma. □

We now discuss the inversion of the Stieltjes transforms $X_\eta(x)$, $\eta=0, 1$, and find the measures with respect to which our polynomials are orthogonal. Let

$$X_\eta(x)=\int_0^\infty (x-t)^{-1}\,d\psi(t; \eta). \tag{3.9}$$

We can easily see from (3.5) that

$$2\pi i\psi'(t; \eta)=X_\eta(t-i0^+)-X_\eta(t+i0^+). \tag{3.10}$$

It readily follows from (3.10) that ψ' vanishes identically on $(-\infty, \gamma^2)$ because $X_\eta(x)$ is single-valued in the complex plane cut along $(-\infty, \gamma^2)$. Clearly,

$$\sqrt{\gamma^2-t\pm i0^+}=\pm i\sqrt{t-\gamma^2},\quad t\in[\gamma^2, \infty). \tag{3.11}$$

Define

$$D=-i\sqrt{t-\gamma^2}+\tfrac{1}{2}(1+a+b+\beta-\alpha), \qquad E=i\sqrt{t-\gamma^2}+\tfrac{1}{2}(1+a+b+\beta-\alpha). \tag{3.12}$$

It is easy to see from (3.4), (3.11), and (3.12) that

$$X_0(t-i0)-X_0(t+i0)$$

$$=\frac{{}_3F_2\left(\begin{matrix}D-a, D-b, D-\alpha-1\\ 1-2i\sqrt{t-\gamma^2}, D\end{matrix}\middle|1\right)}{\alpha(D-1){}_3F_2\left(\begin{matrix}D-a, D-b, D-\alpha-1\\ 1-2i\sqrt{t-\gamma^2}, D-1\end{matrix}\middle|1\right)}-\text{complex conjugate}. \tag{3.13}$$

We then apply (4.4.3) in Slater [17] and obtain

$$\alpha[t-\gamma^2+\{(a+b+\beta-\alpha-1)/2\}^2]\{t-\gamma^2\}^{-1/2}\{X_0(t-i0)-X_0(t+i0)\}$$

$$=\lim_{z\to 1^-}\frac{(1-z)^{a+b+\beta+1-D-E}{}_3F_2\left(\begin{matrix}a, b, \beta+1\\ D, E\end{matrix}\middle|z\right)}{\left|{}_3F_2\left(\begin{matrix}D-a, D-b, D-\alpha-1\\ 1-2i\sqrt{t-\gamma^2}, D-1\end{matrix}\middle|z\right)\right|^2}.$$

Lemma 2 shows that the limit of the numerator on the right-hand side of the above formula is $\Gamma(E)\Gamma(D)\Gamma(a+b+\beta+1-D-E)/\{\Gamma(1+\beta)\Gamma(a)\Gamma(b)\}$. This fact, when combined with (3.10) establishes, after some simplification, the following result.

THEOREM 3. *The weight function $\psi'(t;0)$ is supported on $[\gamma^2,\infty)$ where it is given by*

$$\psi'(t;0)=\frac{\Gamma(\alpha)\sqrt{t-\gamma^2}}{\pi\alpha\Gamma(a)\Gamma(b)\Gamma(\beta+1)} \tag{3.14}$$

$$\cdot\left|\frac{\Gamma(i\sqrt{t-\gamma^2}+\frac{1}{2}(a+b-1+\beta-\alpha))}{{}_3F_2\left(\begin{matrix}i\sqrt{t-\gamma^2}+\frac{1}{2}(a-b+1+\beta-\alpha), i\sqrt{t-\gamma^2}+\frac{1}{2}(b-a+1+\beta-\alpha), i\sqrt{t-\gamma^2}+\frac{1}{2}(a+b-1-\alpha-\beta),\\ 1+2i\sqrt{t-\gamma^2}, i\sqrt{t-\gamma^2}+\frac{1}{2}(a+b+\beta-\alpha-1)\end{matrix}\middle|1\right)}\right|^2\cdot$$

It is clear from (1.1), (1.2), and (1.6) that the dual Hahn polynomials remain invariant if we interchange a and b, or interchange α and β. Therefore $\psi'(t,\eta)$ must also be invariant under the same operations. The right-hand side of formula (3.14) is symmetric in a and b but does not exhibit the symmetry of $\psi'(t;0)$ in α and β. The more symmetric form

$$\pi\Gamma(a)\Gamma(b)\Gamma(\alpha+1)\Gamma(\beta+1)\psi'(t;0)$$

$$=\sqrt{t-\gamma^2}\left|\frac{\Gamma\{\frac{1}{2}(3+\alpha+\beta-a-b)+i\sqrt{t-\gamma^2}\}\Gamma\{\frac{1}{2}(a+b+\beta-\alpha-1)+i\sqrt{t-\gamma^2}\}}{\Gamma(1+2i\sqrt{t-\gamma^2})}\right|^2 \tag{3.15}$$

$$\cdot\left|{}_3F_2\left(\begin{matrix}a-1, & b-1, & \frac{1}{2}(1+\alpha+\beta-a-b)+i\sqrt{t-\gamma}\\ \frac{1}{2}(a+b+\alpha-\beta-1)+i\sqrt{t-\gamma^2}, & \frac{1}{2}(a+b+\beta-\alpha-1)+i\sqrt{t-\gamma^2}\end{matrix}\middle|1\right)\right|^2,$$

for $t\in[\gamma^2,\infty)$ can be obtained from the transformation

$${}_3F_2\left(\begin{matrix}\mathbf{A}, \mathbf{B}, \mathbf{C}\\ \mathbf{D}, \mathbf{E}\end{matrix}\middle|1\right)=\frac{\Gamma(\mathbf{E})\Gamma(\mathbf{S})}{\Gamma(\mathbf{E}-\mathbf{C})\Gamma(\mathbf{S}+\mathbf{C})}{}_3F_2\left(\begin{matrix}\mathbf{D}-\mathbf{A}, \mathbf{D}-\mathbf{B}, \mathbf{C}\\ \mathbf{D}, \mathbf{S}+\mathbf{C}\end{matrix}\middle|1\right),$$

with $\mathbf{S}=\mathbf{D}+\mathbf{E}-\mathbf{A}-\mathbf{B}-\mathbf{C}$.

One can treat the case $\eta=1$ similarly, that is $\mu_0=0$. In this case we need the following lemma.

LEMMA 4. *The ${}_3F_2$'s satisfy the contiguous relation*

$$Z\,{}_3F_2\left(\begin{matrix}X, Y, Z+1\\ U, V\end{matrix}\middle|1\right)=(V-1)\,{}_3F_2\left(\begin{matrix}X, Y, Z\\ U, V-1\end{matrix}\middle|1\right)+(1+Z-V)\,{}_3F_2\left(\begin{matrix}X, Y, Z\\ U, V\end{matrix}\middle|1\right). \tag{3.16}$$

Proof. This lemma follows from

$$Z(Z+1)_n=(Z)_n(Z+n)=(Z)_n\{(Z-V+1)+(V+n-1)\}. \qquad \square$$

Remark. Lemma 4 is known but is not easily accessible.

Using the notation

$$X=\tfrac{1}{2}(a-b+\beta-\alpha+1)-i\sqrt{t-\gamma^2}, \qquad Y=\tfrac{1}{2}(1+b-a+\beta-\alpha)-i\sqrt{t-\gamma^2},$$

$$U=1-2i\sqrt{t-\gamma^2},$$

$$V=\tfrac{1}{2}(1+a+b-\beta-\alpha)-i\sqrt{t-\gamma^2}, \qquad Z=\tfrac{1}{2}(a+b-1-\beta-\alpha)-i\sqrt{t-\gamma^2},$$

together with (3.4), (3.9), and (3.10) we find

$$\begin{aligned}
&\alpha\left|\gamma-i\sqrt{t-\gamma^2}\right|^2\{X_1(t+i0)-X_1(t-i0)\}\left|{}_3F_2\left(\begin{matrix}X, Y, Z+1\\ U, V\end{matrix}\middle|1\right)\right|^2\\
&\quad=Z\,{}_3F_2\left(\begin{matrix}X, Y, Z+1\\ U, V\end{matrix}\middle|1\right){}_3F_2\left(\begin{matrix}\bar X, \bar Y, \bar Z\\ \bar U, \bar V\end{matrix}\middle|1\right)-\text{complex conjugate}\\
&\quad=(V-1)\,{}_3F_2\left(\begin{matrix}X, Y, Z\\ U, V-1\end{matrix}\middle|1\right){}_3F_2\left(\begin{matrix}\bar X, \bar Y, \bar Z\\ \bar U, \bar V\end{matrix}\middle|1\right)\\
&\qquad+(1+Z-V)\left|{}_3F_2\left(\begin{matrix}X, Y, Z\\ U, V\end{matrix}\middle|1\right)\right|^2-\text{complex conjugate},
\end{aligned}$$

which, since $1+Z-V$ is real, simplifies to

$$(V-1)\,{}_3F_2\left(\begin{matrix}X, Y, Z\\ U, V-1\end{matrix}\middle|1\right){}_3F_2\left(\begin{matrix}\bar X, \bar Y, \bar Z\\ \bar U, \bar V\end{matrix}\middle|1\right)-\text{complex conjugate}.$$

The above quantity appeared earlier in the inversion of $X_0(x)$ and was simplified using (4.4.3) in Slater [15]. This enables us to find the absolutely continuous component of the measure $d\psi(t;1)$. When we combine this result with (3.14) we obtain the following theorem.

THEOREM 5. *Let $\gamma=(1+\alpha+\beta-a-b)/2$ (as in (2.4)). The weight functions $\psi'(t;\eta)$ when $\eta=0,1$, are given by*

$$\begin{aligned}
\psi'(t;\eta)=&\frac{t^{-\eta}\sqrt{t-\gamma^2}\left|\Gamma\{1-\eta+\gamma+i\sqrt{t-\gamma^2}\}\right|^2}{\pi\Gamma(a)\Gamma(b)\Gamma(\alpha+1)\Gamma(\beta+1)\left|\Gamma\{1+2i\sqrt{t-\gamma^2}\}\right|^2}\\
&\cdot\left|\frac{\Gamma\{\eta+\alpha-\gamma-i\sqrt{t-\gamma^2}\}\Gamma\{\eta+\beta-\gamma+i\sqrt{t-\gamma^2}\}}{{}_3F_2\left(\begin{matrix}a+\eta-1, b+\eta-1, \eta-\gamma+i\sqrt{t-\gamma^2}\\ \eta+\alpha-\gamma+i\sqrt{t-\gamma^2}, \eta+\beta-\gamma+i\sqrt{t-\gamma^2}\end{matrix}\middle|1\right)}\right|^2
\end{aligned} \tag{3.17}$$

valid for $t\in(\gamma^2,\infty)$.

The isolated jumps of ψ coincide with poles of its Stieltjes transform. We believe, but have been unable to prove, that the functions appearing in the denominators of $X_\eta(x)$, $\eta=0,1$ do not vanish for real values of x if α and β are positive when $\mu_0=\alpha\beta$, or $\alpha>-1$ and $\beta>-1$ when $\mu_0=0$.

4. Explicit and transformation formulas. We first derive the explicit formula

$$P_n(x;a,b,\alpha,\beta,\eta)=\sum_{j=0}^{n}\frac{(\gamma+a-i\sqrt{x-\gamma^2})_j(\gamma+b-i\sqrt{x-\gamma^2})_j(-\gamma+i\sqrt{x-\gamma^2})_{n-j}}{(n-j)!(\alpha+1)_j(\beta+1)_j}$$

$$\cdot\,{}_3F_2\left(\begin{matrix}\alpha,\ \beta,\ \gamma-\eta-i\sqrt{x-\gamma^2}\\ \gamma+a-i\sqrt{x-\gamma^2},\ \gamma+b-i\sqrt{x-\gamma^2}\end{matrix}\middle|1\right). \tag{4.1}$$

Proof of (4.1). We make the substitution

$$G(x,w)=w^{-\beta}(1-w)^{\gamma-i\sqrt{x-\gamma^2}}F(x,w) \tag{4.2}$$

in (2.2) and find that the function $F(x,w)$ satisfies the differential equation

$$w(1-w)F''+[C-(1+A+B)w]F'-ABF=\alpha\beta w^{\beta-1}(1-w)^{\eta-\gamma+i\sqrt{x-\gamma^2}}, \tag{4.3}$$

where $'$ denotes differentiation with respect to w. We then solve (4.3) using the Frobenius method. Clearly,

$$F(x,w)=\sum_{0}^{\infty}c_n w^{n+\beta},$$

with $c_0=1$. Equating coefficients of various powers of w leads to the two-term recurrence relations

$$(\alpha+n+1)(\beta+n+1)c_{n+1}-(n+a+\gamma-i\sqrt{x-\gamma^2})$$

$$\cdot\,(n+b+\gamma-i\sqrt{x-\gamma^2})c_n=\frac{\alpha\beta(\gamma-\eta-i\sqrt{x-\gamma^2})_{n+1}}{(n+1)!}$$

whose solution, subject to the initial condition $c_o=1$, is

$$c_n=\frac{(a+\gamma-i\sqrt{x-\gamma^2})_n(b+\gamma-i\sqrt{x-\gamma^2})_n}{(\alpha+1)_n(\beta+1)_n}\sum_{k=0}^{n}\frac{(\alpha)_k(\beta)_k(\gamma-\eta-i\sqrt{x-\gamma^2})_k}{k!(a+\gamma-i\sqrt{x-\gamma^2})_k(b+\gamma-i\sqrt{x-\gamma^2})_k}.$$

This and (4.2) establish (4.1) and the proof is complete. □

Observe that when α or β vanishes our polynomials reduce to the continuous dual Hahn polynomials and (4.1) provides a representation of continuous dual Hahn polynomials as multiples of ${}_3F_2$'s. Note also that (4.1) remains valid if we reverse the signs of all square roots involved because the left-hand side is a real polynomial in x.

An interesting representation of the associated continuous dual Hahn polynomials is discovered if we change variables in (2.2) as follows. Set

$$G(x,w)=w^{-\beta}(1-w)^{\beta-a}P(x,z), \qquad \text{with } w=z/(1+z). \tag{4.4}$$

It is readily seen that $P(x,z)$ satisfies the differential equation

$$z(1+z)\frac{\partial^2 P}{\partial z^2}+[1+\alpha-\beta+(2+a-b+\alpha-\beta)z]\frac{\partial P}{\partial z}$$

$$+[(a-\beta)(a-b+1)+x]P=\alpha\beta z^{\beta-1}(1+z)^{1-\alpha-\eta}.$$

We again apply the Frobenius method and let

$$P(x,z)=\sum_{n=0}^{\infty}h_n z^{n+\beta}.$$

We then find that the h_n's satisfy the two-term recurrence relation

$$(n+\beta+1)(n+\alpha+1)h_{n+1}$$

$$+\{(n+\beta)(n+a-b+\alpha+1)+(a-\beta)(\alpha-b+1)+x\}h_n=\alpha\beta(-1)^{n+1}\frac{(a+\eta-1)_{n+1}}{(n+1)!},$$

whose solution subject to $h_0=1$ is

$$h_n=(-1)^n\frac{(a+\gamma+i\sqrt{x-\gamma^2})_n(a+\gamma-i\sqrt{x-\gamma^2})_n}{(\alpha+1)_n(\beta+1)_n}$$

$$\cdot\sum_{k=0}^{n}\frac{(\alpha)_k(\beta)_k(a+\eta-1)_k}{k!(a+\gamma+i\sqrt{x-\gamma^2})_k(a+\gamma-i\sqrt{x-\gamma^2})_k}.$$

Thus we have

$$G(x,w)=(1-w)^{-a}\sum_{k=0}^{\infty}h_k\left(\frac{w}{1-w}\right)^k=\sum_{j,k=0}^{\infty}\frac{(a+k)_j}{j!}h_kw^{k+j}.$$

Replacing $(a+k)_j$ by $(a)_{k+j}/(a)_k$ leads to the explicit formula

$$P_n(x;a,b,\alpha,\beta,\eta)=\frac{(a)_n}{n!}\sum_{k=0}^{n}\frac{(-n)_k(a+\gamma+i\sqrt{x-\gamma^2})_k(a+\gamma-i\sqrt{x-\gamma^2})_k}{(a)_k(\alpha+1)_k(\beta+1)_k}$$

$$\cdot\sum_{j=0}^{k}\frac{(\alpha)_j(\beta)_j(a+\eta-1)_j}{j!(a+\gamma+i\sqrt{x-\gamma^2})_j(a+\gamma-i\sqrt{x-\gamma^2})_j}. \tag{4.5}$$

When α or β vanishes, the right-hand side of (4.5) reduces to the familiar ${}_3F_2$ representation of continuous dual Hahn polynomials.

Observe that the right-hand side of (4.5) is obviously invariant under interchanging α and β. On the other hand, the λ_n's and μ_n's in (1.1) and (1.2) are symmetric in a and b (see (1.6)). Therefore the polynomials P_n must also be symmetric in a and b.

THEOREM 6. *The right-hand side of* (4.5) *is a symmetric function of a and b where* γ *is as in* (2.4).

Theorem 6 is a generalization of the Whipple transformation [17] to a double series.

Acknowledgment. We thank the referees for their help in improving the presentation of some of our results, for correcting several typographical errors, and for pointing out a minor error in the original version of this paper.

REFERENCES

[1] G. ANDREWS AND R. ASKEY, *Classical orthogonal polynomials*, in Polynomes Orthogonaux et Applications, Lecture Notes in Mathematics 1171, Springer-Verlag, Berlin, New York, 1985, pp. 36-82.

[2] R. ASKEY AND M. E. H. ISMAIL, *Recurrence relations, continued fractions and orthogonal polynomials*, Mem. Amer. Math. Soc., 300 (1984), pp. 1-108.

[3] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985), pp. 1-54.

[4] R. ASKEY AND J. WIMP, *Associated Laguerre and Hermite polynomials*, Proc. Roy. Soc. Edinburgh, Sect. A, 96 (1984), pp. 15-37.

[5] N. T. BAILEY, *The Elements of Stochastic Processes*, John Wiley, New York, 1964.

[6] A. ERDELYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[7] M. E. H. ISMAIL, J. LETESSIER, AND G. VALENT, *Linear birth and death models and associated Laguerre polynomials*, J. Approx. Theory, 55 (1988).

[8] S. KARLIN AND J. MCGREGOR, *The differential equations of birth and death processes, and the Stieltjes moment problem*, Transactions Amer. Math. Soc., 85 (1958), pp. 489-546.

[9] ———, *Linear growth, birth and death processes*, J. Math. Mech. (now Indiana Math. J.), 7 (1958), pp. 643-662.

[10] J. LABELLE, *Tableau d'Askey*, University of Quebec, Montreal, Canada, 1984.

[11] J. LETESSIER AND G. VALENT, *The generating function method for quadratic asymptotically symmetric birth and death processes*, SIAM J. Appl. Math., 44 (1983), pp. 773-783.

[12] ———, *Dual birth and death processes and orthogonal polynomials*, SIAM J. Appl. Math., 46 (1986), pp. 393-405.

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

[14] F. POLLACZEK, *Sur une généralisation des polynomes de Jacobi*, Mémor. Sci. Math., 131 (1956).

[15] E. D. RAINVILLE, *Special Functions*, Chelsea, New York, 1971.

[16] J. A. SHOHAT AND J. D. TAMARKIN, *The Problem of Moments*, revised edition, Mathematical Surveys, Vol. 1, American Mathematical Society, Providence, RI, 1950.

[17] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.

[18] G. SZEGO, *Orthogonal Polynomials*, Fourth Edition, Vol. 23, Colloquium Publications, American Mathematical Society, Providence, RI, 1975.

[19] J. WIMP, *Explicit formulas for the associated Jacobi polynomials and some applications*, Canadian J. Math., 39 (1987), pp. 983-1000.

[20] D. MASSON, *The rotating harmonic oscillator eigenvalue problem. I. Continued fractions and analytic continuations*, J. Math. Phys., 24 (1983), pp. 2074-2088.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 738-743, May 1989

016

ON THE ZEROS OF POLYNOMIALS ORTHOGONAL ON THE SEMICIRCLE*

WALTER GAUTSCHI†

Abstract. It is shown that the polynomials $\pi_n(\cdot; w)$ orthogonal in the sense of [W. Gautschi, H. J. Landau, and G. V. Milovanović, *Constr. Approx.*, 3 (1987), pp. 389-404] on the unit upper semicircle need not necessarily have all their zeros in the interior of the unit upper semidisc, not even for weight functions w that are symmetric, $w(-z)=w(z)$. A symmetric weight function w_a (depending on a parameter a) is exhibited, which has the property that $\pi_n(\cdot; w_a)$ for any fixed even n has a zero on the imaginary axis with imaginary part greater than one, provided a is large enough. Similarly, a weight function w^a is constructed for which the analogous property holds for $\pi_n(\cdot; w^a)$, n odd.

Key words. complex orthogonal polynomials, indefinite inner product, zeros

AMS(MOS) subject classifications. 30C10, 30C15, 33A65

1. In [3], [4] we introduced polynomials that are orthogonal on the semicircle with respect to the (non-Hermitian) inner product

$$(p, q)=\int_0^{\pi} p(e^{i\theta})q(e^{i\theta})w(e^{i\theta})\,d\theta. \tag{1.1}$$

Here, w is a "weight function" analytic on the semidisc $D_+=\{z\in\mathbb{C}: |z|<1, \operatorname{Im} z>0\}$, nonnegative on $(-1, 1)$ and integrable over ∂D_+. We have shown that under the assumption

$$\operatorname{Re}\int_0^{\pi} w(e^{i\theta})\,d\theta\neq 0 \tag{1.2}$$

there exists a unique system $\{\pi_n\}_{n=0}^{\infty}$ of monic polynomials $\pi_n(\cdot)=\pi_n(\cdot; w)$ such that

$$\deg \pi_n=n, \quad n=0, 1, 2, \cdots, \qquad (\pi_k, \pi_l)\begin{cases}=0, & k\neq l,\\ \neq 0, & k=l.\end{cases} \tag{1.3}$$

They possess many of the properties familiar from orthogonal polynomials on the real line, such as satisfying a three-term recurrence relation and a second-order linear differential equation (for special weight functions), and in fact can be expressed as (complex) linear combinations of two successive polynomials orthogonal on the interval $(-1, 1)$ with respect to the same weight function w. They give rise to Gauss-type quadrature rules for integration over the semicircle and to new, possibly more stable, quadrature formulae for evaluating Cauchy principal value integrals (see [3, §§ 7, 8]). Since the nodes of these quadrature rules involve the zeros of the polynomials π_n in (1.3), a study of the qualitative properties of these zeros is of interest.

In [4] we have shown that for weight functions analytic in $D=\{z\in\mathbb{C}: |z|<1\}$, symmetric in the sense

$$w(-z)=w(z) \quad \text{for all } z\in D, \tag{1.4}$$

and satisfying

$$w(x)\geqq 0 \quad \text{on } (-1, 1), \quad w(0)>0, \tag{1.5}$$

* Received by the editors December 8, 1987; accepted for publication (in revised form) August 1, 1988. This research was supported in part by National Science Foundation grant CCR-8704404.

† Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907.

all zeros of π_n are contained in D_+ with the possible exception of a single (simple) zero iy, $y \geqq 1$. For the Gegenbauer weight $w(z) = (1-z^2)^{\lambda-1/2}$, the exceptional case can only arise if $n = 1$ and $-\frac{1}{2} < \lambda \leqq 0$. Likewise, no exceptional cases seem to occur for Jacobi weights $w(z) = (1-z)^\alpha (1+z)^\beta$, $\alpha > -1$, $\beta > -1$, if $n \geqq 2$, as was observed by numerical computation. We might be led to believe that this absence of exceptional cases prevails for arbitrary weight functions w. In this note we show, however, that this is not so, not even for symmetric weight functions. We exhibit symmetric functions w for which $\pi_n(\cdot; w)$, for arbitrary fixed n, has a zero iy with $y \geqq 1$.

2. Let $b_k = b_k(w)$, $k = 1, 2, 3, \cdots$, be the coefficients in the recurrence formula

$$y_{k+1} = xy_k - b_k y_{k-1}, \quad k = 0, 1, 2, \cdots, \qquad y_{-1} = 0, \quad y_0 = 1 \tag{2.1}$$

satisfied by the polynomials $p_n(x; w)$ orthogonal on the interval $(-1, 1)$ relative to the symmetric weight function w. We recall from the proof of Theorem 6.5 and equations (5.2), (5.4) of [4] that iy is a zero of $\pi_n(\cdot; w)$ if and only if

$$\omega_n(y) - \theta_{n-1} = 0, \tag{2.2}$$

where

$$\omega_1(y) = y, \quad \omega_k(y) = y + \frac{b_{k-1}}{\omega_{k-1}(y)}, \quad k = 2, 3, \cdots, \tag{2.3}$$

$$\theta_{n-1} = \begin{cases} \dfrac{b_1 b_3 \cdots b_{n-1}}{b_2 b_4 \cdots b_{n-2}} \dfrac{\pi}{m_0}, & n \text{ even}, \\ \dfrac{b_2 b_4 \cdots b_{n-1}}{b_1 b_3 \cdots b_{n-2}} \dfrac{m_0}{\pi}, & n \text{ odd}, \end{cases} \tag{2.4}$$

and

$$m_0 = \int_{-1}^{1} w(x)\, dx = 2\int_0^1 w(x)\, dx, \tag{2.5}$$

the weight function w having been normalized to satisfy

$$w(0) = 1. \tag{2.6}$$

If $n = 1$ or $n = 2$, empty products in (2.4) are assumed to be one. Equation (2.2) holds for some $y \geqq 1$ if and only if

$$\omega_n(1) - \theta_{n-1} \leqq 0. \tag{2.7}$$

Indeed, since $\omega_n(y) \to \infty$ as $y \to \infty$, inequality (2.7) trivially implies (2.2) for some $y \geqq 1$. Conversely, if (2.2) holds for some $y \geqq 1$, but (2.7) (if $n \geqq 2$) does not, the left-hand side of (2.2), hence $\pi_n(iy)$, would have either two distinct zeros >1, or a double zero >1, which is impossible by Theorem 6.2 of [4]. By (2.3), we can write (2.7) in the form

$$1 + \frac{b_{n-1}}{1+}\,\frac{b_{n-2}}{1+} \cdots \frac{b_1}{1} \leqq \theta_{n-1}. \tag{2.8}$$

We now show that (2.8), for any fixed $n \geqq 1$, can always be achieved for some suitable weight function w.

3. It is necessary to distinguish the cases n even and n odd. In the former case, (2.8) becomes

$$1 + \frac{b_{n-1}}{1+}\,\frac{b_{n-2}}{1+} \cdots \frac{b_1}{1} \leqq \frac{b_1 b_3 \cdots b_{n-1}}{b_2 b_4 \cdots b_{n-2}} \frac{\pi}{m_0}. \tag{3.1}$$

It is clear that we can enforce (3.1) to hold if we can find a family of weight functions w for which m_0 tends to zero and the b_k remain bounded and bounded away from zero. Such a family of weight functions (keeping in mind that they should be analytic in D, satisfy (1.4), and be normalized by (2.6)) is given by

$$w(z) = w_a(z) = \frac{1+\sqrt{a/\pi}\, e^{-az^2}}{1+\sqrt{a/\pi}}, \qquad a>0. \tag{3.2}$$

The fact that w_a also satisfies (1.2) follows from Theorem 5.1 of [4]. We note that the second term in the numerator of (3.2), for real $z=x$, is an approximation to the Dirac delta function $\delta(x)$, to which it converges as $a\to\infty$. It follows that, for any polynomial p,

$$\begin{aligned}\left(1+\sqrt{\frac{a}{\pi}}\right)\int_{-1}^{1} w_a(x)p(x)\,dx &\to \int_{-1}^{1}[1+\delta(x)]p(x)\,dx \\ &= \int_{-1}^{1} p(x)\,dx + p(0) \quad \text{as } a\to\infty.\end{aligned} \tag{3.3}$$

In particular, putting $p(x)\equiv 1$,

$$m_0 \sim 3\sqrt{\pi/a}, \qquad a\to\infty. \tag{3.4}$$

Furthermore,

$$\lim_{a\to\infty} b_k(w_a) = b_{k,\infty} > 0, \quad \cdot\ k=1,2,3,\cdots, \tag{3.5}$$

where $b_{k,\infty}$ are the recursion coefficients of the monic polynomials orthogonal with respect to the weight function $1+\delta(x)$ on $[-1,1]$ (Legendre weight plus Dirac function centered at the origin). It follows from (3.4) and (3.5) that for a sufficiently large, (3.1) will be true (even with strict inequality). The proof of (3.5) is deferred to § 4.

Assume next that n is odd. Then, (2.8) becomes

$$1+\frac{b_{n-1}}{1+}\,\frac{b_{n-2}}{1+}\cdots\frac{b_1}{1} \leqq \frac{b_2b_4\cdots b_{n-1}}{b_1b_3\cdots b_{n-2}}\,\frac{m_0}{\pi}. \tag{3.6}$$

We now want m_0 to be large and may choose, for example,

$$w(z) = w^a(z) = 1+az^2, \qquad a>0. \tag{3.7}$$

Then

$$m_0 = \tfrac{2}{3}a+2 \tag{3.8}$$

and

$$\lim_{a\to\infty} b_k(w^a) = b_k^{\infty} > 0, \qquad k=1,2,3,\cdots, \tag{3.9}$$

where b_k^{∞} are the recursion coefficients of the monic polynomials orthogonal with respect to the weight function x^2 on $[-1,1]$. Again, from (3.8) and (3.9) it follows that (3.6) will be true for a sufficiently large. It remains to prove (3.5) and (3.9).

4. We denote the moments of w by m_k,

$$m_{2r+1} = 0, \qquad m_{2r} = 2\int_0^1 x^{2r}w(x)\,dx > 0. \tag{4.1}$$

The recursion coefficients $b_k(w)$ can be expressed in terms of Hankel determinants:

$$\Delta_n(m) = \det\,(m_{i+j})_{\substack{i=0,1,\cdots,n-1\\ j=0,1,\cdots,n-1}}, \qquad \Delta_0 = 1, \tag{4.2}$$

by means of [1, p. 19]

$$b_k(w)=\frac{\Delta_{k-1}(m)\Delta_{k+1}(m)}{[\Delta_k(m)]^2}, \qquad k=1,2,3,\cdots. \tag{4.3}$$

In the case of $w(x)=w_a(x)$ [cf. (3.2)], we have by (3.3)

$$m_r\sim\sqrt{\pi/a}\,m_{r,\infty}, \qquad a\to\infty, \quad r=0,1,2,\cdots, \tag{4.4}$$

where $m_{r,\infty}$ are the moments of the weight function $1+\delta(x)$ on $[-1,1]$. Therefore,

$$\Delta_n(m)\sim\left(\frac{\pi}{a}\right)^{n/2}\Delta_n(m_\infty), \qquad a\to\infty,$$

and, consequently, by (4.3),

$$b_k(w_a)\sim\frac{\Delta_{k-1}(m_\infty)\Delta_{k+1}(m_\infty)}{[\Delta_k(m_\infty)]^2}, \qquad a\to\infty,$$

that is,

$$b_k(w_a)\to b_{k,\infty} \quad \text{as } a\to\infty. \tag{4.5}$$

Likewise, for $w(x)=w^a(x)$ [cf. (3.7)],

$$m_r\sim am_r^\infty, \quad a\to\infty, \quad r=0,1,2,\cdots,$$

where m_r^∞ are the moments of the weight function x^2 on $[-1,1]$, and thus,

$$\Delta_n(m)\sim a^n\Delta_n(m^\infty), \qquad a\to\infty,$$

giving

$$b_k(w^a)\to b_k^\infty \quad \text{as } a\to\infty. \tag{4.6}$$

This proves the assertions in (3.5) and (3.9).

We remark that instead of one in (2.7) we could have selected any number larger than one, which means that the zeros of $\pi_n(\cdot\,;w_a)$ and $\pi_n(\cdot\,;w^a)$ on the imaginary axis can be made to have arbitrarily large imaginary parts by choosing a sufficiently large.

5. We now confirm the validity of the construction in § 3 numerically by computing the zeros of $\pi_n(\cdot\,;w_a)$, n even, and of $\pi_n(\cdot\,;w^a)$, n odd, for the critical value $a=a_n^*$ (which should yield a zero at i) and a few selected values $a>a_n^*$. We compute these zeros in terms of eigenvalues of a real tridiagonal (nonsymmetric) matrix, as indicated in [4, § 6.1], the coefficients $b_k(w_a)$ and $b_k(w^a)$ being generated by the "discretized Stieltjes procedure" (cf. [2, § 2.2]).

Table 5.1 shows the values of a_n^* for $n=2(1)10$ obtained to eight significant decimal digits by using the bisection method on (2.2) where $y=1$. The zeros of $\pi_n(\cdot\,;w_a)$

TABLE 5.1
Values of a_n^* for $n=2(1)10$.

n	a_n^*	n	a_n^*
2	55.274946	3	17.009652
4	250.25427	5	46.413430
6	798.58573	7	89.537192
8	1951.2926	9	146.34390
10	4037.4957		

TABLE 5.2
Zeros of $\pi_n(\cdot\,; w_a)$, $a=(1+\kappa)a_n^*$, $\kappa=0, \frac{1}{2}, 1, \infty$, *where* $n=2(2)10$.

n	κ	Zeros					
2	0	$.225i$	$1.000i$				
	.5	$.177i$	$1.264i$				
	1.0	$.152i$	$1.472i$				
	∞	0	∞i				
4	0	$.065i$	$1.000i$	$\pm.797+.038i$			
	.5	$.053i$	$1.237i$	$\pm.791+.035i$			
	1.0	$.046i$	$1.435i$	$\pm.788+.032i$			
	∞	0	∞i	$\pm.775$			
6	0	$.031i$	$1.000i$	$\pm.912+.011i$	$\pm.559+.051i$		
	.5	$.025i$	$1.251i$	$\pm.911+.010i$	$\pm.553+.044i$		
	1.0	$.022i$	$1.461i$	$\pm.910+.009i$	$\pm.550+.039i$		
	∞	0	∞i	$\pm.906$	$\pm.538$		
8	0	$.018i$	$1.000i$	$\pm.951+.004i$	$\pm.751+.022i$	$\pm.421+.046i$	
	.5	$.015i$	$1.262i$	$\pm.951+.004i$	$\pm.749+.019i$	$\pm.416+.039i$	
	1.0	$.013i$	$1.479i$	$\pm.950+.004i$	$\pm.747+.017i$	$\pm.413+.034i$	
	∞	0	∞i	$\pm.949$	$\pm.742$	$\pm.406$	
10	0	$.012i$	$1.000i$	$\pm.969+.002i$	$\pm.841+.011i$	$\pm.623+.026i$	$\pm.335+.040i$
	.5	$.010i$	$1.269i$	$\pm.969+.002i$	$\pm.840+.010i$	$\pm.620+.022i$	$\pm.331+.033i$
	1.0	$.008i$	$1.492i$	$\pm.969+.002i$	$\pm.839+.009i$	$\pm.618+.019i$	$\pm.330+.029i$
	∞	0	∞i	$\pm.968$	$\pm.836$	$\pm.613$	$\pm.324$

TABLE 5.3
Zeros of $\pi_n(\cdot\,; w^a)$, $a=(1+\kappa)a_n^*$, $\kappa=0, \frac{1}{2}, 1, \infty$, *where* $n=3(2)9$.

n	κ	Zeros				
3	0	$1.000i$	$\pm.781+.046i$			
	.5	$1.356i$	$\pm.776+.038i$			
	1.0	$1.710i$	$\pm.774+.032i$			
	∞	∞i	$\pm.775$			
5	0	$1.000i$	$\pm.909+.012i$	$\pm.541+.057i$		
	.5	$1.407i$	$\pm.908+.010i$	$\pm.536+.044i$		
	1.0	$1.807i$	$\pm.907+.008i$	$\pm.535+.035i$		
	∞	∞i	$\pm.906$	$\pm.538$		
7	0	$1.000i$	$\pm.951+.005i$	$\pm.747+.023i$	$\pm.405+.051i$	
	.5	$1.429i$	$\pm.950+.004i$	$\pm.744+.018i$	$\pm.402+.038i$	
	1.0	$1.847i$	$\pm.950+.003i$	$\pm.743+.015i$	$\pm.401+.030i$	
	∞	∞i	$\pm.949$	$\pm.742$	$\pm.406$	
9	0	$1.000i$	$\pm.969+.002i$	$\pm.839+.012i$	$\pm.618+.027i$	$\pm.321+.044i$
	.5	$1.440i$	$\pm.969+.002i$	$\pm.838+.009i$	$\pm.615+.020i$	$\pm.320+.032i$
	1.0	$1.868i$	$\pm.968+.002i$	$\pm.837+.008i$	$\pm.614+.016i$	$\pm.320+.025i$
	∞	∞i	$\pm.968$	$\pm.836$	$\pm.613$	$\pm.324$

and $\pi_n(\cdot; w^a)$ for $n = 2(2)10$ and $n = 3(2)9$, respectively, where $a = (1+\kappa)a_n^*$, $\kappa = 0$, $\frac{1}{2}$, 1, ∞, are listed in Tables 5.2 and 5.3. Although they were computed to eight significant digits, only three-digit values are shown because of space considerations. If $a \to \infty$ (i.e., $\kappa \to \infty$), it follows from $\pi_n = p_n - i\theta_{n-1}p_{n-1}$ (cf. [4, eq. (2.9)]) and $\theta_{n-1} \to \infty$ that the (finite) zeros of π_n tend to those of p_{n-1}, the orthogonal polynomial of degree $n-1$ relative to the limiting weight function $w_\infty(x) = 1 + \delta(x)$ and $w^\infty(x) = x^2$, for n even and odd, respectively. These limiting weights are not as unrelated as we might think at first. We have, in fact,

$$p_{n-1}(x; w_\infty) = xp_{n-2}(x; w^\infty), \qquad n(\text{even}) \geqq 2,$$

since each side is easily seen to be orthogonal on $[-1, 1]$ to all powers of degree $\leqq n-2$ with respect to the constant weight function $w \equiv 1$, and hence equal to the monic Legendre polynomial of degree $n-1$. This is why the limiting zeros for $\kappa = \infty$ in Table 5.2 and Table 5.3 are the same.

It can be seen that for n even, there are two zeros on the imaginary axis moving in opposite directions as a increases from a_n^* to ∞, one up from i to $i\infty$ (cf. the remark at the end of § 4), the other down from some iy_n^*, $0 < y_n^* < 1$, to zero. For n odd, there is one zero on the imaginary axis moving up from i to $i\infty$.

It is also easy to compute the coefficients $b_{k,\infty}$ and b_k^∞ and to observe numerically the convergence in (3.5) and (3.9).

REFERENCES

[1] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[2] W. GAUTSCHI, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 289-317.

[3] W. GAUTSCHI AND G. V. MILOVANOVIĆ, *Polynomials orthogonal on the semicircle*, J. Approx. Theory, 46 (1986), pp. 230-250.

[4] W. GAUTSCHI, H. J. LANDAU, AND G. V. MILOVANOVIĆ, *Polynomials orthogonal on the semicircle*, II, Constr. Approx., 3 (1987), pp. 389-404.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 744-760, May 1989

017

UNIFORM ASYMPTOTIC EXPANSIONS FOR WHITTAKER'S CONFLUENT HYPERGEOMETRIC FUNCTIONS*

T. M. DUNSTER†

Abstract. The asymptotic behavior, as $\kappa \to \infty$, of the Whittaker confluent hypergeometric functions $M_{\kappa,\mu}(z)$ and $W_{\kappa,\mu}(z)$ is examined. Asymptotic expansions are derived in terms of Bessel and Airy functions, the results being uniformly valid for real values of κ and μ such that $0 \leqq \mu/\kappa \leqq A < 1$ (A an arbitrary constant), and for all complex values of the argument z. Explicit error bounds are available for all the approximations.

Key words. confluent hypergeometric functions, asymptotic expansions

AMS(MOS) subject classifications. 33A30, 34E05

1. Introduction. In this paper we aim to derive asymptotic expansions for solutions of Whittaker's equation

$$\frac{d^2w}{dz^2} = \left\{\frac{1}{4} - \frac{\kappa}{z} + \frac{\mu^2 - \frac{1}{4}}{z^2}\right\} w, \tag{1.1}$$

this being a form of the confluent hypergeometric equation (see, for example, Olver [5, p. 260]).

We consider the case where the parameters κ and μ are real, with κ large and μ either large or small, and the independent variable z complex with unbounded absolute value.

Previously there have been a large number of investigations into the asymptotic behavior of solutions of Whittaker's equation, and for a comprehensive survey of the results prior to 1975 the reader is referred to Olver [6, pp. 127-132]. In this survey Olver remarks that there are several outstanding problems to be tackled. One of these is the case of $|\mu| \to \infty$ with $1 - \delta \leqq \kappa/\mu \leqq 1 + \delta$; (here and throughout δ is used generically as an arbitrary small positive constant). This problem has since been successfully treated by Olver [8] who has derived asymptotic solutions in terms of parabolic cylinder functions. The results are uniformly valid for κ, μ, and z either all real or all purely imaginary, and are furnished with explicit error bounds. The asymptotic solutions were derived by an application of the asymptotic theory of second-order differential equations having coalescing turning points (Olver [7]).

A second outstanding problem is the case where $\kappa \to \infty$ with $-1 + \delta \leqq \mu/\kappa \leqq 1 - \delta$, and this is the problem that we address in the present paper. To get a better insight into this particular case let us reformulate (1.1) into the form

$$\frac{d^2w}{d\xi^2} = \left\{\frac{\kappa^2(\xi - \xi_1(\alpha^2))(\xi - \xi_2(\alpha^2))}{4\xi^2} - \frac{1}{4\xi^2}\right\} w, \tag{1.2}$$

where for convenience we write

$$\xi = z/\kappa, \tag{1.3}$$

$$\alpha = 2\mu/\kappa, \tag{1.4}$$

* Received by the editors January 27, 1988; accepted for publication (in revised form) July 1, 1988.

† Department of Mathematics, University of British Columbia, British Columbia, V6T 1Y4, Canada. Present address, Department of Mathematical Sciences, San Diego State University, San Diego, California 92182-0314.

and

$$\xi_1 = 2-(4-\alpha^2)^{1/2}, \tag{1.5a}$$

$$\xi_2 = 2+(4-\alpha^2)^{1/2}. \tag{1.5b}$$

When κ is large, (1.2) is characterized by having a regular singularity at $\xi=0$, an irregular singularity at infinity, and turning points at $\xi=\xi_1(\alpha^2)$ and $\xi=\xi_2(\alpha^2)$. The position of these turning points depends on the value of α^2. When κ and μ are real we have the following four cases:

(i) $\alpha^2=0$; ξ_1 coalesces with the singularity $\xi=0$, and $\xi_2=4$.

(ii) $0<\alpha^2<4$; ξ_1 lies in the interval $(0,2)$, and ξ_2 lies in the interval $(2,4)$.

(iii) $\alpha^2=4$; ξ_1 and ξ_2 coalesce at the point $\xi=2$.

(iv) $4<\alpha^2<\infty$; ξ_1 and ξ_2 no longer lie on the real axis, and are complex conjugates.

Our results then will be uniformly valid for $0\leqq\alpha^2\leqq 4-\delta$, or in other words for ξ_1 either coalescing with the pole at the origin or taking positive real values. As $\alpha^2\to 4$ the turning points coalesce and the results of Olver [8] are applicable; our results can be regarded as complementary to Olver's.

The plan of the paper is as follows. In § 2 we present definitions and relevant properties of the Whittaker functions that are to be approximated. Also, we record connection formulae that will be used relating to these functions.

In §§ 3 and 4 we transform Whittaker's equation (1.2) into two different forms. In the first of these new equations the coefficient of κ^2 has a double pole and a simple zero which coalesces with the pole as $\alpha\to 0$. The coefficient of κ^2 in the second equation has only one transition point, a simple zero.

In § 5 we apply the general theory of a coalescing turning point and singularity (Boyd and Dunster [2]). This general theory is applicable to the first of the transformed equations, and provides asymptotic expansions for its solutions in terms of Bessel functions. These approximations for Whittaker functions are uniformly valid in certain regions of the complex ξ plane which include both $\xi=0$ and $\xi=\xi_1$, but not $\xi=\xi_2$.

We then apply the general theory of a simple turning point in the complex plane to the second of the transformed equations. This theory is given in Olver [5, Chap. 11]. The resulting asymptotic expansions involve Airy functions and are uniformly valid in certain regions of the complex ξ plane that include $\xi=\xi_2$, but not $\xi=\xi_1$. The domains of validity for the Bessel function and Airy function expansions overlap and together cover the entire complex ξ plane. The principal results are summarized in § 6.

Of the previous investigations into Whittaker functions with $|\kappa|$ large we mention two special cases of our results. The first is the work of Erdélyi and Swanson [3] who construct asymptotic approximations for $M_{\kappa,\mu}(\kappa\xi)$ and $W_{\kappa,\mu}(\kappa\xi)$ in terms of Bessel and Airy functions. These results are valid for *fixed* nonnegative μ, i.e., $\alpha=O(\kappa^{-1})$. Accordingly, the investigation is that of the case of a simple pole and one fixed turning point. The resulting approximations are uniformly valid for $|\kappa|$ large (real or complex) and ξ real with $0\leqq\xi<\infty$. Skovgaard [9] has extended these results to asymptotic expansions with ξ complex.

Neither Erdélyi and Swanson, nor Skovgaard, supply error bounds for their approximations; however, for the real variable case, Olver [5, pp. 412-413, 446-447] provides error bounds for both the Airy and Bessel function expansions.

More recently, Baumgartner [1] has investigated the Whittaker function $M_{\kappa,\mu}(\kappa\xi)$ in great detail, and has derived a uniform approximation in terms of the Bessel function $J_\nu(z)$, which is uniformly valid for positive μ and κ with $0\leqq\mu/\kappa\leqq 1-\delta$, and for ξ lying in the finite interval $0\leqq\xi\leqq\xi_2-\delta$. A very thorough error analysis is given and explicit error bounds are derived. Our results are considerably more general than those

of Baumgartner: we derive full expansions rather than the leading term; we approximate in turn the Whittaker functions $M_{\kappa,\mu}(\kappa\xi)$, $W_{\kappa,\mu}(\kappa\xi)$, $W_{-\kappa,\mu}(\kappa\xi e^{-\pi i})$, and $W_{-\kappa,\mu}(\kappa\xi e^{\pi i})$; and the results are uniformly valid in domains that cover the entire complex ξ plane. These expansions are uniformly valid for real nonnegative values of μ and κ such that $0\leqq\mu/\kappa\leqq 1-\delta$, and are readily extended to negative values of μ and κ by means of appropriate connection formulae.

2. Whittaker functions: Definitions, characteristic properties and connection formulae.

2.1. Standard solutions. The Whittaker functions we approximate are $M_{\kappa,\mu}(z)$, $W_{\kappa,\mu}(z)$, $W_{-\kappa,\mu}(z e^{\pi i})$, and $W_{-\kappa,\mu}(z e^{-\pi i})$, each being a solution of (1.1). These functions are defined by

$$M_{\kappa,\mu}(z)=e^{-z/2}z^{1/2+\mu}\,{}_1F_1(\tfrac{1}{2}+\mu-\kappa;\,1+2\mu;\,z), \tag{2.1}$$

and

$$W_{\kappa,\mu}(z)=\frac{\Gamma(-2\mu)}{\Gamma(\frac{1}{2}-\mu-\kappa)}M_{\kappa,\mu}(z)+\frac{\Gamma(2\mu)}{\Gamma(\frac{1}{2}+\mu-\kappa)}M_{\kappa,-\mu}(z), \tag{2.2}$$

where ${}_1F_1$ denotes the confluent hypergeometric function. The limiting form of (2.2) is taken when 2μ is an integer. The four solutions are linearly independent of one another for all nonnegative values of μ and κ, the only exception being that $M_{\kappa,\mu}(z)$ and $W_{k,\mu}(z)$ are multiples of one another when $\kappa-\mu-\frac{1}{2}$ is a nonnegative integer.

2.2. Behavior at singular points. Whittaker's equation has a regular singularity at $z=0$ and an irregular singularity at infinity. The behavior of the four solutions at these singular points is given as follows (with κ and μ assumed fixed):

$$M_{\kappa,\mu}(z)=z^{\mu+1/2}\{1+O(z)\}\quad\text{as } z\to 0, \tag{2.3}$$

$$\begin{aligned}W_{\kappa,\mu}(z)=&\frac{\Gamma(-2\mu)}{\Gamma(\frac{1}{2}-\mu-\kappa)}z^{\mu+1/2}\{1+O(z)\}\\&+\frac{\Gamma(2\mu)}{\Gamma(\frac{1}{2}+\mu-\kappa)}z^{-\mu+1/2}\{1+O(z)\}\quad\text{as } z\to 0,\end{aligned} \tag{2.4}$$

$$\begin{aligned}M_{\kappa,\mu}(z)=&\frac{\Gamma(2\mu+1)}{\Gamma(\mu+\kappa+\frac{1}{2})}e^{(\mu-\kappa+1/2)\pi i}z^{\kappa}e^{-z/2}\{1+O(z^{-1})\}\\&+\frac{\Gamma(2\mu+1)}{\Gamma(\mu-\kappa+\frac{1}{2})}z^{-\kappa}e^{z/2}\{1+O(z^{-1})\}\\&\qquad\text{as } z\to\infty,\quad -\pi/2<\arg z<3\pi/2,\end{aligned} \tag{2.5}$$

$$W_{\kappa,\mu}(z)=z^{\kappa}e^{-z/2}\{1+O(z^{-1})\}\quad\text{as } z\to\infty,\quad -3\pi/2<\arg z<3\pi/2, \tag{2.6}$$

$$W_{-\kappa,\mu}(z e^{\pi i})=e^{-\kappa\pi i}z^{-\kappa}e^{z/2}\{1+O(z^{-1})\}\quad\text{as } z\to\infty,\quad -5\pi/2<\arg z<\pi/2, \tag{2.7}$$

$$W_{-\kappa,\mu}(z e^{-\pi i})=e^{\kappa\pi i}z^{-\kappa}e^{z/2}\{1+O(z^{-1})\}\quad\text{as } z\to\infty,\quad -\pi/2<\arg z<5\pi/2. \tag{2.8}$$

From (2.3), (2.6), (2.7), and (2.8) we perceive that the characterizing properties of the four functions are that, for nonnegative μ, $M_{\kappa,\mu}(z)$ is recessive at the origin, and that for all values of μ and κ the functions $W_{\kappa,\mu}(z)$, $W_{-\kappa,\mu}(z e^{\pi i})$, and $W_{-\kappa,\mu}(z e^{-\pi i})$ are recessive at infinity in the sectors $-\pi/2<\arg z<\pi/2$, $-3\pi/2<\arg z<-\pi/2$, and $\pi/2<\arg z<3\pi/2$, respectively.

2.3. Connection formulae. We will require the following connection formulae:

$$W_{\kappa,\mu}(z)=\frac{1}{2\pi}\Gamma\left(\kappa+\mu+\frac{1}{2}\right)\Gamma\left(\kappa-\mu+\frac{1}{2}\right)\cdot[e^{(\kappa-1/2)\pi i}W_{-\kappa,\mu}(z\,e^{\pi i})+e^{-(\kappa-1/2)\pi i}W_{-\kappa,\mu}(z\,e^{-\pi i})], \tag{2.9}$$

$$M_{\kappa,\mu}(z)=\frac{1}{2\pi}\Gamma(2\mu+1)\Gamma\left(\kappa-\mu+\frac{1}{2}\right)[e^{\mu\pi i}W_{-\kappa,\mu}(z\,e^{\pi i})+e^{-\mu\pi i}W_{-\kappa,\mu}(z\,e^{-\pi i})]. \tag{2.10}$$

Our results are valid for $\mu\geqq 0$ and $\kappa>0$. We can extend these parameters to negative values using (2.2) and the following relations:

$$M_{-\kappa,\mu}(z\,e^{\pm\pi i})=e^{\pm(\mu+1/2)\pi i}M_{\kappa,\mu}(z), \tag{2.11}$$

$$W_{\kappa,-\mu}(z)=W_{\kappa,\mu}(z). \tag{2.12}$$

3. Domains containing the transition points $\xi=0$ and $\xi=\xi_1$: Preliminary transformations. In §§ 3 and 4 our purpose is to transform Whittaker's equation (1.1) to two new forms. The first, given here, resembles Bessel's equation and has a finite regular singularity. The second, given in § 4, resembles Airy's equation. Both have simple turning points. Having made these transformations, we can derive asymptotic solutions and then identify them with standard Whittaker functions.

Then, using a Liouville transformation that transforms both the dependent and independent variables, we transform (1.1) to the Bessel equation form. Following [2, eq. (2.1)], we define a new independent variable $\zeta(\xi)$ and dependent variable $W(\zeta)$ by

$$\frac{d\zeta}{d\xi}=\frac{\zeta}{\xi}\left(\frac{\xi_1-\xi}{\alpha^2-\zeta}\right)^{1/2}(\xi_2-\xi)^{1/2}, \tag{3.1}$$

$$W(\zeta)=\left(\frac{d\zeta}{d\xi}\right)^{1/2}w(\xi), \tag{3.2}$$

where the branches of the fractional powers will be specified shortly. With these transformations, Whittaker's equation (1.1) is transformed to the equation

$$\frac{d^2W}{d\zeta^2}=\left\{\kappa^2\left[\frac{\alpha^2}{4\zeta^2}-\frac{1}{4\zeta}-\frac{1}{4\zeta^2}\right]+\frac{\psi(\alpha,\zeta)}{\zeta}\right\}W, \tag{3.3}$$

where

$$\frac{\psi(\alpha,\zeta)}{\zeta}=\xi'^{1/2}\frac{d^2}{d\zeta^2}(\xi'^{-1/2})-\frac{\xi'^2}{4\xi^2}+\frac{1}{4\zeta^2}; \tag{3.4}$$

primes ($'$) denote differentiation with respect to ζ. An explicit expression for ψ can be obtained by using (3.2); a straightforward calculation yields

$$\psi(\alpha,\zeta)=\frac{1}{16(\alpha^2-\zeta)^2}\left[\zeta+4\alpha^2-\frac{4(\alpha^2-\zeta)^3}{[(\xi_1-\xi)(\xi_2-\xi)]^3}\left(\frac{\xi}{\zeta}\right)(\xi^3-16\xi^2+4(1-\alpha^3)\xi+4\alpha^2)\right]. \tag{3.5}$$

We now examine the $\xi-\zeta$ transformation in more detail. First, integration of (3.1) gives the relationship

$$\int_{\alpha^2}^{\zeta}\frac{(\alpha^2-t)^{1/2}}{2t}\,dt=\int_{\xi_1}^{\xi}\frac{(\xi_1-t)^{1/2}(\xi_2-t)^{1/2}}{2t}\,dt. \tag{3.6}$$

The lower integration limits are chosen so that the turning point $\xi = \xi_1$ of the original equation (1.1) is mapped to the turning point $\zeta = \alpha^2$ of the transformed equation (3.3). Note also that both integrands have a singularity at the origin, and as a consequence the pole $\xi = 0$ of (1.1) corresponds to the pole $\zeta = 0$ of (3.3).

The relationship (3.6) as it stands is not well defined as there are branch points at the transition points $\xi = 0$, ξ_1, ξ_2, and $\zeta = 0$ and α^2. We therefore introduce branches as follows. First, with regard to the poles at the origin, we temporarily introduce a branch cut along the negative real axis in both cases, and assign the principal values of the arguments for both ξ and ζ. Throughout this paper we assume that ξ (and ζ) take their principal values; appropriate connection formulae can be used to extend all subsequent results to other ranges of arg ξ.

The branches of the fractional powers are chosen as follows. The integrand on the left-hand side of (3.6) is assumed to be negative imaginary just above the semi-infinite interval (α^2, ∞), positive imaginary below the same interval, and continuous elsewhere. The integrand on the right-hand side is assumed to be negative imaginary just above the interval (ξ_1, ξ_2), positive imaginary below the interval, and continuous elsewhere.

The integrals in (3.6) can be evaluated explicitly (see, for example, Gradshteyn and Ryzhik, [4, pp. 83, 84]). We obtain

$$
\begin{aligned}
&(\alpha^2-\zeta)^{1/2}-\frac{\alpha}{2}\ln\left\{\frac{\alpha+(\alpha^2-\zeta)^{1/2}}{\alpha-(\alpha^2-\zeta)^{1/2}}\right\}\\
&\quad=\frac{1}{2}(\xi_1-\xi)^{1/2}(\xi_2-\xi)^{1/2}+\frac{\alpha}{2}\ln\left\{\frac{\xi}{\xi_1}\,\frac{\alpha^2-2\xi_1}{\alpha^2-2\xi+\alpha(\xi_1-\xi)^{1/2}(\xi_2-\xi)^{1/2}}\right\}\\
&\qquad+\ln\left\{\frac{2-\xi_1}{2-\xi-(\xi_1-\xi)^{1/2}(\xi_2-\xi)^{1/2}}\right\}.
\end{aligned}
\tag{3.7}
$$

Let $\underline{\Delta}$ denote the ζ domain given by $|\arg \zeta| < \pi$ with all points on the interval $[\zeta_2, \infty)$ excluded (ζ_2 denoting $\zeta(\xi_2)$). We will confine our attention to $\underline{\Delta}$ and the corresponding ξ domain Δ: both Δ and $\underline{\Delta}$, together with corresponding points, are illustrated in Figs. 1(a) and 1(b). The curves CD, CD' are given by

$$
\operatorname{Re}\int_{\xi_2}^{\xi}\frac{(\xi_1-t)^{1/2}(\xi_2-t)^{1/2}\,dt}{2t}=0,
$$

and are asymptotic to a line parallel to the imaginary axis.

The mapping $\zeta(\xi)$ is conformal within Δ and hence the inverse $\xi(\zeta)$ is conformal within $\underline{\Delta}$. It follows that $\psi(\alpha, \zeta)$ is holomorphic in $\underline{\Delta}$, and that, moreover, it is uniformly

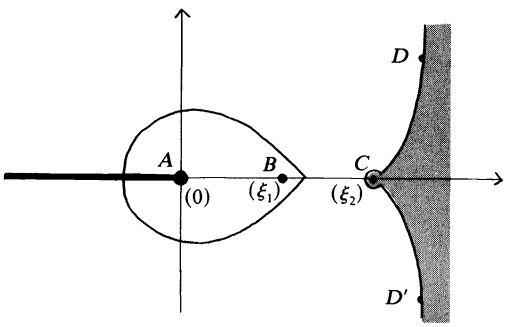


FIG. 1(a). *Domain Δ in ξ plane.*

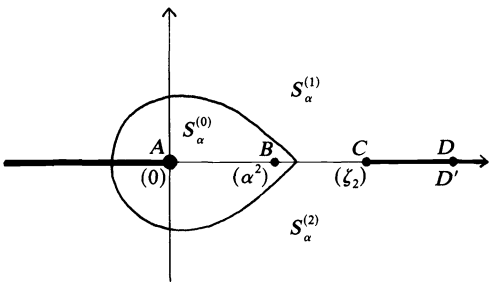


FIG. 1(b). *Domain $\underline{\Delta}$ in ζ plane.*

continuous for $\zeta \in \underline{\Delta}$ and $\alpha \in [0, 2-\delta]$; (see [2, eq. (2.6) and the paragraph after Lemma 1]).

To identify solutions of the transformed equation (3.3) with standard Whittaker functions we will need to know the asymptotic behavior of $\zeta(\xi)$ as $\xi \to 0$, and also as $\xi \to -\infty \pm i0$. (Here and throughout $\pm i0$ indicates a complex number being above or below a cut.)

First, we find from (3.7) that $\zeta \to 0$ as $\xi \to 0$, and more specifically

$$\alpha - \alpha \ln 2\alpha + \frac{\alpha}{2} \ln \zeta + O(\zeta) = \frac{\alpha}{2} + \frac{\alpha}{2} \ln \left\{ \frac{\alpha^2 - 2\xi_1}{2\alpha^2 \xi_1} \right\} + \ln \left\{ \frac{2-\xi_1}{2-\alpha} \right\} + \frac{\alpha}{2} \ln \xi + O(\xi). \tag{3.8}$$

On collecting terms and exponentiating we arrive at

$$\zeta = c(\alpha)\xi + O(\xi^2) \quad \text{as } \xi \to 0, \tag{3.9}$$

where

$$c(\alpha) = \frac{2}{e} \left(\frac{\alpha^2 - 2\xi_1}{\xi_1} \right) \left(\frac{2-\xi_1}{2-\alpha} \right)^{2/\alpha}. \tag{3.10}$$

We remark that (3.9), (3.10) is uniformly valid for $\alpha \in [0, 2-\delta]$; (see [5, eqs. (A3), (A4)]). In particular when $\alpha = 0$ the limiting form of (3.9) applies:

$$\zeta \sim 4\xi \quad \text{as } \xi \to 0. \tag{3.11}$$

Next we examine $\zeta(\xi)$ as $\xi \to -\infty \pm i0$. We find from (3.7) that $\zeta \to -\infty \pm i0$, such that

$$(-\zeta)^{1/2} + O\left(\frac{1}{\zeta}\right) = \frac{1}{2}(-\xi) + \frac{\alpha}{2} \ln \left\{ \frac{\alpha^2 - 2\xi_1}{\xi_1(2+\alpha)} \right\} + \ln \left\{ \frac{2(2-\xi_1)}{4-\alpha^2} \right\} + \ln(-\xi) + O\left(\frac{1}{\xi}\right), \tag{3.12}$$

the result being uniformly valid for $\alpha \in [0, 2-\delta]$.

Finally we must examine the asymptotic behavior of $\psi(\alpha, \zeta)$ and its derivatives as $\zeta \to \infty$ in $\underline{\Delta}$. From (3.7) we have

$$(-\zeta)^{1/2} \sim \tfrac{1}{2}\xi + O(\ln \xi) \quad \text{as } \xi \to \infty \text{ in } \Delta,$$

and hence from (3.5) we have

$$\psi(\alpha, \zeta) \sim \frac{1}{8(-\zeta)^{3/2}} \quad \text{as } \zeta \to \infty \text{ in } \underline{\Delta}, \tag{3.13}$$

uniformly for $\alpha \in [0, 2-\delta]$.

Also, on differentiating (3.5) s times, we find that the sth derivative of ψ

$$\psi^{(s)}(\alpha, \zeta) = O(\zeta^{-3/2-s}) \quad \text{as } \zeta \to \infty \text{ in } \underline{\Delta}, \tag{3.14}$$

uniformly for $\alpha \in [0, 2-\delta]$.

4. Domains containing the transition point $\xi = \xi_2$: Preliminary transformations. In the previous section Whittaker's equation was transformed, via a Liouville transformation, to a new differential equation from which asymptotic solutions are to be obtained. As we have remarked, the transformation is not regular at $\xi = \xi_2$, and the asymptotic solutions that we obtain are not uniformly valid in a neighborhood of $\xi = \xi_2$. The purpose of this section is to use the general theory of [5] to transform Whittaker's equation, via a different Liouville transformation, to a differential equation from which we can obtain asymptotic solutions that are uniformly valid in a neighborhood of $\xi = \xi_2$.

We use the notation of [5] except that each term here is written with a circumflex (^) to avoid a clash of notation with § 3.

The appropriate Liouville transformation is given by the following (see [5, (3.02)]):

$$\frac{d\hat{\zeta}}{d\xi}=\left(\frac{(\xi-\xi_1)(\xi-\xi_2)}{\hat{\zeta}}\right)^{1/2}\frac{1}{2\xi}, \tag{4.1}$$

$$\hat{W}(\hat{\zeta})=\left(\frac{d\hat{\zeta}}{d\xi}\right)^{1/2}w(\xi). \tag{4.2}$$

The effect of the above transformations is to transform (1.1) into

$$\frac{d^2\hat{W}}{d\hat{\zeta}^2}=\{\kappa^2\hat{\zeta}+\hat{\psi}(\alpha,\hat{\zeta})\}\hat{W}, \tag{4.3}$$

where

$$\hat{\psi}(\alpha,\hat{\zeta})=\frac{5}{16\hat{\zeta}^2}-\frac{\hat{\zeta}\xi}{(\xi-\xi_1)^3(\xi-\xi_2)^3}[\xi^3-16\xi^2+4(1-\alpha^2)\xi+4\alpha^2]. \tag{4.4}$$

For the $\xi-\hat{\zeta}$ map to be regular at the transition point $\xi=\xi_2$, this point must correspond to the turning point $\hat{\zeta}=0$ of the transformed equation (4.3). Thus integration of (4.1) yields

$$\frac{2}{3}\hat{\zeta}^{3/2}=\int_{\xi_2}^{\xi}\frac{[(t-\xi_1)(t-\xi_2)]^{1/2}}{2t}\,dt. \tag{4.5}$$

With regard to the logarithmic singularity at $\xi=0$ we introduce a branch cut along the negative real axis and restrict our attention to $|\arg\xi|<\pi$. We also introduce a cut along the negative real $\hat{\zeta}$ axis and a cut along the real ξ axis from ξ_1 to ξ_2; with these cuts both sides of (4.5) are assumed to be positive for $\hat{\zeta}\in(0,\infty)$ and $\xi\in(\xi_2,\infty)$ and to be continuous elsewhere in the cut plane.

The integral in (4.5) is expressible in terms of elementary functions: we obtain the relationship (cf. (3.7))

$$\begin{aligned}\frac{2}{3}\hat{\zeta}^{3/2}=&\frac{1}{2}(\xi-\xi_1)^{1/2}(\xi-\xi_2)^{1/2}+\frac{\alpha}{2}\ln\frac{\xi}{\xi_2}\left\{\frac{2\xi_2-\alpha^2}{2\xi-\alpha^2-\alpha(\xi-\xi_1)^{1/2}(\xi-\xi_2)^{1/2}}\right\}\\&+\ln\left\{\frac{\xi_2-2}{\xi-2+(\xi-\xi_1)^{1/2}(\xi-\xi_2)^{1/2}}\right\}.\end{aligned} \tag{4.6}$$

The map of the half space $0\leqq\arg\xi<\pi$ is shown in Figs. 2(a) and 2(b) for $\alpha\neq 0$. The map of the half space $-\pi<\arg\xi\leqq 0$ is the conjugate of the $\hat{\zeta}$ domain indicated in Fig. 2(b).

It is possible to obtain asymptotic solutions of (4.3) that are uniformly valid as $\hat{\zeta}\to\infty$ with $\arg\hat{\zeta}=\pm 4\pi/3$, or equivalently as $\xi\to 0\pm i0$. However, these solutions would not hold uniformly at $\xi=0$ when $\alpha\to 0$.

On the other hand, the asymptotic solutions of (3.3) *are* uniformly valid in a neighborhood of $\xi=0$ for $\alpha\in[0,2-\delta]$. Therefore for simplicity we will restrict our attention to a ξ domain that includes $\xi=\xi_2$, but excludes both $\xi=0$ and $\xi=\xi_1$.

To this end we define $\hat{\Delta}$ to be the ξ domain $|\arg\xi|<\pi$ lying outside the pear-shaped region BCB', with a neighborhood of $\xi=\xi_1$ excluded. When $\alpha=0$ $\hat{\Delta}$ is the ξ domain $|\arg\xi|<\pi$ with a neighborhood of $\xi=0$ excluded. The domain $\hat{\Delta}$ and the corresponding $\hat{\zeta}$ domain $\hat{\underline{\Delta}}$, are shown in Figs. 3(a) and 3(b).

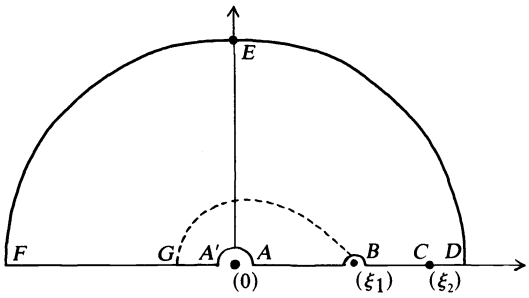


FIG. 2(a). *ξ plane.*

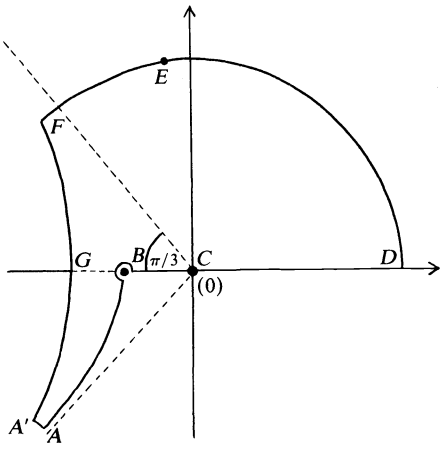


FIG. 2(b). *ζ̂ plane.*

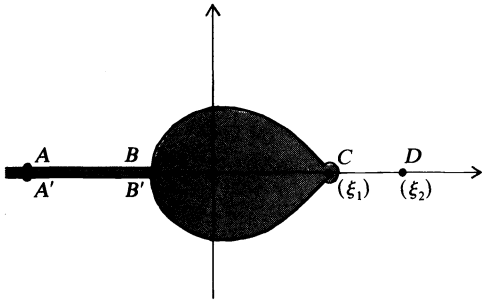


FIG 3(a). *Domain* $\hat{\Delta}$ *in* ξ *plane.*

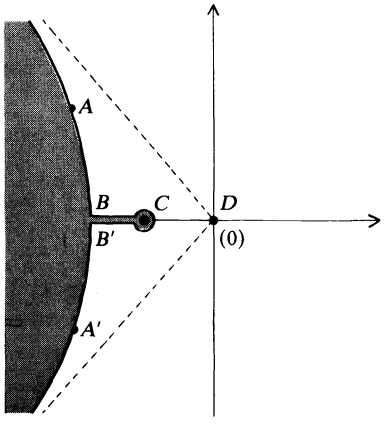


FIG. 3(b). *Domain* $\hat{\underline{\Delta}}$ *in* $\hat{\zeta}$ *plane.*

Since both $\xi = 0$ and $\xi = \xi_1$ do not lie in $\hat{\Delta}$, it follows that $\hat{\zeta}(\xi)$ is holomorphic within $\hat{\Delta}$. Furthermore, since $d\hat{\zeta}/d\xi$ does not vanish in this domain the inverse mapping $\xi(\hat{\zeta})$ is holomorphic within $\hat{\underline{\Delta}}$. We also note that $\hat{\psi}(\alpha, \hat{\zeta})$ is holomorphic within $\hat{\underline{\Delta}}$, as well as being uniformly continuous for $\hat{\zeta} \in \hat{\underline{\Delta}}$ and $\alpha \in [0, 2-\delta]$.

We now record the asymptotic behavior of $\hat{\zeta}(\xi)$ as $\xi \to \infty$. From (4.6) we find that

$$\frac{2}{3}\hat{\zeta}^{3/2} = \frac{1}{2}\xi - \ln \xi + \frac{\alpha}{2}\ln\left\{\frac{2\xi_2 - \alpha^2}{\xi_2(2-\alpha)}\right\} + \ln\left(\frac{1}{2}\xi_2 - 1\right) + O(\xi^{-1}), \qquad |\arg \hat{\zeta}| \leqq \frac{2\pi}{3}. \tag{4.7}$$

In general,

$$\frac{2}{3}\hat{\zeta}^{3/2} = \frac{1}{2}\xi + O(\ln \xi) \quad \text{as } \xi \to \infty \text{ in } \hat{\Delta}, \tag{4.8}$$

and therefore from (4.4) we deduce that

$$\hat{\psi}(\alpha, \hat{\zeta}) \sim -\frac{1}{4\hat{\zeta}^2}, \tag{4.9}$$

and

$$\hat{\psi}^{(s)}(\alpha, \hat{\zeta}) = O(\hat{\zeta}^{-2-s}) \quad \text{as } \hat{\zeta} \to \infty \text{ in } \hat{\underline{\Delta}}. \tag{4.10}$$

5. Asymptotic expansions for Whittaker functions. Having made the preliminary transformations, we now apply the general theories of [2] and [5] to obtain asymptotic solutions of the transformed equations (3.3) and (4.3).

First let us apply the theory of § 5 of [2] to the differential equation (3.3). From Theorem 3 we deduce that, for each value of κ and α and nonnegative integer n, the following solutions of (3.3) exist that are holomorphic in $\underline{\Delta}$:

$$W^{(j)}_{2n+1}(k,\alpha,\zeta)=\zeta^{1/2}\mathscr{C}^{(j)}_{\kappa\alpha}(\kappa\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{\kappa^{2s}} + \frac{\zeta}{\kappa}\,\mathscr{C}^{(j)'}_{\kappa\alpha}(\kappa\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{\kappa^{2s}}+\varepsilon^{(j)}_{2n+1}(\kappa,\alpha,\zeta) \tag{5.1}$$

$$(j=0,1,2),$$

with $\mathscr{C}^{(j)}_{\nu}(z)$ denoting the Bessel function $2J_\nu(z)$, $H^{(1)}_\nu(z)$, $H^{(2)}_\nu(z)$ for $j=0,1,2$, respectively. The coefficients A_s and B_s are given recursively by the integral equations

$$B_s(\alpha,\zeta)=(\alpha^2-\zeta)^{-1/2}\int_{\zeta}^{\alpha^2}(\alpha^2-t)^{-1/2}(tA''_s(\alpha,t)+A'_s(\alpha,t)-\psi(\alpha,t)A_s(\alpha,t))\,dt \tag{5.2}$$

$$(s=0,1,2,\cdots),$$

$$A_s(\alpha,\zeta)=-\zeta B'_{s-1}+\int^{\zeta}\psi(\alpha,t)B_{s-1}(\alpha,t)\,dt+\lambda_s \qquad (s=1,2,\cdots), \tag{5.3a}$$

$$A_0(\alpha,\zeta)=\lambda_0. \tag{5.3b}$$

The numbers λ_s $(s=0,1,2,\cdots)$ are arbitrary constants of integration, and the branches in (5.2) are defined as in § 3. Each of the coefficients is holomorphic in $\underline{\Delta}$, and furthermore they are uniformly continuous for $\zeta\in\underline{\Delta}$ and $\alpha\in[0,2-\delta]$ (provided of course that each constant λ_s is taken to be a continuous function of α).

Before assigning values to the constants λ_s we must record the asymptotic behavior of $A_s(\alpha,\zeta)$ and $B_s(\alpha,\zeta)$ as $\zeta\to\infty$ in $\underline{\Delta}$. From (3.13), (3.14), and Ritt's theorem (see, for example, Olver ([5, pp. 9, 10)]) we can easily establish by induction from (5.2), (5.3a, b) that

$$A_s(\alpha,\zeta)=a_s+O(\zeta^{-1/2}) \qquad (s=1,2,\cdots), \tag{5.4a}$$

$$B_s(\alpha,\zeta)=b_s(-\zeta)^{-1/2}+O(\zeta^{-3/2}) \qquad (s=0,1,2,\cdots), \tag{5.4b}$$

as $\zeta\to\infty$ with $|\arg(-\zeta)|\leqq\pi-\delta$. Here a_s and b_s are numbers independent of ζ. Similarly, we can prove the important result that the variations $\mathscr{V}\{(\zeta-\alpha^2)^{1/2}B_s(\zeta)\}$ converge as $\zeta\to\infty$ with $|\arg(-\zeta)|\leqq\pi-\delta$.

We choose two different sets of values for the constants as follows. For the solutions $W^{(1)}_{2n+1}$ and $W^{(2)}_{2n+1}$ we assign λ_0 (and hence A_0) the value 1, and each of the subsequent constants λ_s so that

$$a_s=0 \qquad (s=1,2,\cdots). \tag{5.5}$$

For each solution $W^{(0)}_{2n+1}$ we again choose $A_0=1$, but now choose each λ_s in turn so that

$$A_s(\alpha,0)=0 \qquad (s=1,2,\cdots). \tag{5.6}$$

Bounds for the error terms $\varepsilon^{(j)}_{2n+1}$ are furnished by Theorem 3 of [2] with u replaced by κ. For the bounds to be meaningful a reference point $\zeta^{(j)}$ in $\underline{\Delta}$ must be assigned for each of the three functions $\varepsilon^{(j)}_{2n+1}$. We take $\zeta^{(0)}=0$, $\zeta^{(1)}=-\infty+i0$, and $\zeta^{(2)}=-\infty-i0$.

Given these reference points we define domains $\underline{\Delta}^{(j)}$, $(j=0, 1, 2)$, to be the set of points in $\underline{\Delta}$ that can be linked to $\zeta^{(j)}$ by a progressive path $\mathscr{P}^{(j)}$ (see [2, p. 439]). Under these circumstances $\underline{\Delta}^{(1)}=\underline{\Delta}^{(2)}=\underline{\Delta}$, and $\underline{\Delta}^{(0)}$ consists of all points in $\underline{\Delta}$ except those lying on the cut on the real axis from $\zeta=\zeta_2$ to $\zeta=\infty$.

The regions $S_\alpha^{(j)}$, $(j=0, 1, 2)$, which are defined in [2], § 5, are illustrated in Fig. 1(b). Each of the solutions $W_{2n+1}^{(j)}$ is characteristically recessive within $S_\alpha^{(j)}$ and dominant elsewhere in $\underline{\Delta}^{(j)}$.

The bounds for $\varepsilon_{2n+1}^{(j)}$ given in Theorem 3 of [2] imply, for each $j=0$, 1, 2, that the ratio $\varepsilon_{2n+1}^{(j)}(\kappa, \alpha, \zeta)[\zeta^{1/2}\mathscr{C}_{\kappa\alpha}^{(j)}(\kappa\zeta^{1/2})]^{-1}$ is $O(\kappa^{-2n-1})$ uniformly for $\alpha\in[0, 2-\delta]$, and for all points ζ in $\underline{\Delta}^{(j)}$ except those near the zeros of the denominator. Furthermore, the same ratio vanishes as $\zeta\to\zeta^{(j)}$. These two important observations are deduced from the properties of the auxiliary functions E and M given in § 5 of [2].

Let us now identify the standard Whittaker functions with the asymptotic solutions (5.1). First we note that $(d\zeta/d\xi)^{1/2}M_{\kappa,\mu}(\kappa\xi)$, regarded as a function of ζ, and $W_{2n+1}^{(0)}(\kappa, \alpha, \zeta)$ are solutions of (3.3) that are recessive at the origin. It follows that, to within a multiplicative constant, the two solutions must be identical. Therefore there exists a constant $c_{2n+1}^{(0)}$ such that

$$M_{\kappa,\mu}(\kappa\xi)=c_{2n+1}^{(0)}\left(\frac{\alpha^2-\zeta}{\xi_1-\xi}\right)^{1/4}(\xi_2-\xi)^{-1/4}\left(\frac{\xi}{\zeta}\right)^{1/2}W_{2n+1}^{(0)}(\kappa, \alpha, \zeta). \tag{5.7}$$

The constant can be determined by comparing both sides of (5.7) at some particular value of ζ. It is convenient to do so at $\zeta=0$ because, as noted above, the ratio $\varepsilon_{2n+1}^{(0)}(\kappa, \alpha, \zeta)[\zeta^{1/2}2J_{\kappa\alpha}(\kappa\zeta^{1/2})]^{-1}$ vanishes as $\zeta\to 0$.

From (2.3) and (3.9) we find that the left-hand side of (5.7) is equal to

$$(\kappa/c(\alpha))^{\mu+1/2}\zeta^{\mu+1/2}(1+O(\zeta)) \quad \text{as } \zeta\to 0, \tag{5.8}$$

where, for the moment, α and κ are assumed to be nonzero and held fixed.

Next, on using the known behavior of $J_\nu(z)$ and $J_\nu'(z)$ near $z=0$ (see Olver [5, p. 436]), and on referring to (3.9), (5.1), and (5.6) we find that the right-hand side of (5.7) is equal to

$$c_{2n+1}^{(0)}\left(\frac{\alpha^2}{\xi_1}\right)^{1/4}\xi_2^{-1/4}c(\alpha)^{-1/2}\frac{2(\kappa/2)^{\kappa\alpha}}{\Gamma(\kappa\alpha+1)}\left\{1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha, 0)}{\kappa^{2s+1}}\right\}\zeta^{(\kappa\alpha+1)/2}(1+O(\zeta)) \tag{5.9}$$

as $\zeta\to 0$.

On equating (5.8) and (5.9), and invoking (1.4) and (3.10), we arrive at the desired relation:

$$c_{2n+1}^{(0)}=\frac{\kappa^{1/2}}{2}\Gamma(\kappa\alpha+1)\left(\frac{2\xi_1 e}{\kappa(\alpha^2-2\xi_1)}\right)^{\kappa\alpha/2}\left(\frac{2-\alpha}{2-\xi_1}\right)^{\kappa}\left\{1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha, 0)}{\kappa^{2s+1}}\right\}^{-1}. \tag{5.10}$$

The assumption that κ and α be held fixed can now be relaxed, and an asymptotic expansion for $M_{\kappa,\mu}(\kappa\xi)$, uniformly valid for $\zeta\in\underline{\Delta}^{(0)}$ and $\alpha\in[0, 2-\delta]$, is now given by (5.7) and (5.10). This and all subsequent expansions will be expressed again in terms of the original variables in § 6.

Next, consider the asymptotic solution $W_{2n+1}^{(1)}(\kappa, \alpha, \zeta)$. This function shares the property with the Whittaker function $(d\zeta/d\xi)^{1/2}W_{-\kappa,\mu}(\kappa\xi e^{-\pi i})$ of being a solution of (3.3) that is recessive as $\zeta\to-\infty+i0$. It therefore follows that a constant $c_{2n+1}^{(1)}$ exists such that

$$W_{-\kappa,\mu}(\kappa\xi e^{-\pi i})=c_{2n+1}^{(1)}\left(\frac{\alpha^2-\zeta}{\xi_1-\xi}\right)^{1/4}(\xi_2-\xi)^{-1/4}\left(\frac{\xi}{\zeta}\right)^{1/2}W_{2n+1}^{(1)}(\kappa, \alpha, \zeta). \tag{5.11}$$

The value of the constant is established similarly to the foregoing analysis, namely when we compare the ratio of both sides of the equation at a convenient value of ζ. In this case we compare both sides of the equation as $\zeta \to -\infty + i0$. Thus we have

$$c_{2n+1}^{(1)} = \lim_{\zeta \to -\infty + i0} \left(\frac{\xi_1 - \xi}{\alpha^2 - \zeta}\right)^{1/4} \left[(\xi_2 - \xi)^{1/4} \left(\frac{\zeta}{\xi}\right)^{1/2} \frac{W_{-\kappa,\mu}(\kappa \xi e^{-\pi i})}{W_{2n+1}^{(1)}(\kappa, \alpha, \zeta)} \right]. \tag{5.12}$$

On employing (2.8), (3.12), (5.1), (5.4a, b), (5.5), together with the known asymptotic behavior of $H_\nu^{(1)}(z)$ and $H_\nu^{(1)'}(z)$ at infinity (see Olver [5, p. 238]), we can evaluate the above limit. After some straightforward calculation we find that

$$c_{2n+1}^{(1)} = -\left(\frac{\pi\kappa}{2}\right)^{1/2} e^{i\kappa\alpha\pi/2} \left(\frac{2(2-\xi_1)}{\kappa(4-\alpha^2)}\right)^{\kappa} \left(\frac{\alpha^2 - 2\xi_1}{\xi_1(2+\alpha)}\right)^{\kappa\alpha/2} \left\{1 - \sum_{s=0}^{n-1} \frac{b_s}{\kappa^{2s+1}}\right\}^{-1}. \tag{5.13}$$

This expression together with (5.11) gives an asymptotic expansion for $W_{-\kappa,\mu}(\kappa\xi e^{-\pi i})$ that is uniformly valid for $\zeta \in \underline{\Delta}$ and $\alpha \in [0, 2-\delta]$. The identification of $W_{2n+1}^{(2)}$ is similar. Both this function and the Whittaker function $(d\zeta/d\xi)^{1/2} W_{-\kappa,\mu}(\kappa\xi e^{\pi i})$ are solutions of (3.3) that are recessive in $S_\alpha^{(2)}$. We deduce the existence of a constant $c_{2n+1}^{(2)}$ such that

$$W_{-\kappa,\mu}(\kappa\xi e^{\pi i}) = c_{2n+1}^{(2)} \left(\frac{\alpha^2 - \zeta}{\xi_1 - \xi}\right)^{1/4} (\xi_2 - \xi)^{-1/4} \left(\frac{\xi}{\zeta}\right)^{1/2} W_{2n+1}^{(2)}(\kappa, \alpha, \zeta), \tag{5.14}$$

the value of the constant being found, similarly to (5.13), to be

$$c_{2n+1}^{(2)} = e^{-i\kappa\alpha\pi} c_{2n+1}^{(1)}. \tag{5.15}$$

Again, these results are uniformly valid for $\zeta \in \underline{\Delta}$, $\alpha \in [0, 2-\delta]$.

Now consider the $\hat{\zeta}$ domain $\hat{\underline{\Delta}}$. Applying Theorem 9.1 of [5] to the transformed equation (4.3) yields asymptotic solutions that are uniformly valid in this region. We obtain the three solutions

$$\hat{W}_{2n+1,j}(\kappa, \hat{\zeta}) = Ai_j(\kappa^{2/3}\hat{\zeta}) \sum_{s=0}^{n} \frac{\hat{A}_s(\hat{\zeta})}{\kappa^{2s}} + \frac{Ai_j'(\kappa^{2/3}\hat{\zeta})}{\kappa^{4/3}} \sum_{s=0}^{n-1} \frac{\hat{B}_s(\hat{\zeta})}{\kappa^{2s}} + \hat{\varepsilon}_{2n+1}(\kappa, \hat{\zeta}) \tag{5.16}$$

$$(j = 0, 1, -1),$$

where $\mathrm{Ai}_j(z)$ denotes the Airy function of complex argument $\mathrm{Ai}(z e^{-2\pi i j/3})$. The coefficients $\hat{A}_s$ and $\hat{B}_s$ are defined recursively by $\hat{A}_0(\hat{\zeta}) = 1$,

$$\hat{B}_s(\hat{\zeta}) = \frac{1}{2} \hat{\zeta}^{-1/2} \int_0^{\hat{\zeta}} \{\hat{\psi}(t)\hat{A}_s''(t) - \hat{A}_s''(t)\} t^{-1/2}\, dt, \tag{5.17}$$

and

$$\hat{A}_{s+1}(\hat{\zeta}) = -\frac{1}{2} \hat{B}_s(\hat{\zeta}) + \int_0^{\hat{\zeta}} \hat{\psi}(t)\hat{B}(t)\, dt + \hat{\lambda}_{s+1} \qquad (s = 0, 1, 2, \cdots). \tag{5.18}$$

Each $\hat{\lambda}_s$ $(s = 1, 2, \cdots)$ is an arbitrary integration constant for which a value will be assigned shortly.

The explicit error bounds given by (9.03) of [5] are meaningful—provided a reference point $\hat{\zeta} = \hat{\alpha}_j$ in $\hat{\underline{\Delta}}$ is assigned for each of the solutions $\hat{W}_{2n+1,j}$ $(j = 0, \pm 1)$. Before we do so we observe that

$$\hat{A}_s(\hat{\zeta}) = \hat{a}_s + O(\hat{\zeta}^{-3/2}), \tag{5.19}$$

$$\hat{B}_s(\hat{\zeta}) = \hat{b}_s \hat{\zeta}^{-1/2} + O(\hat{\zeta}^{-2}), \tag{5.20}$$

as $\hat{\zeta} \to \infty$ in $\hat{\underline{\Delta}}$, where each $\hat{a}_s$ and $\hat{b}_s$ is a constant. Also, we can prove that the variations $\mathcal{V}(\hat{\zeta}^{1/2}\hat{B}_s(\hat{\zeta}))$ converge as $\hat{\zeta} \to \infty$ in $\hat{\underline{\Delta}}$. These results follow from (4.9), (4.10), (5.17),

(5.18), and Ritt's theorem. It is justifiable then to choose the reference points at infinity: we take $\hat{\alpha}_0=\infty$, $\hat{\alpha}_1=\infty\, e^{2\pi i/3}$, and $\hat{\alpha}_{-1}=\infty\, e^{-2\pi i/3}$. With these choices it follows from the error bounds that for each $j(0,\pm 1)$ the ratio $\hat{\varepsilon}_{2n+1,j}(\kappa,\hat{\zeta})/\mathrm{Ai}_j(\kappa^{2/3}\hat{\zeta})$ is $O(\kappa^{-2n-1})$ uniformly for $\hat{\zeta}\in\hat{\underline{\Delta}}$, except near the zeros of the denominator and near the cut interval BCB'. Moreover, the same ratio vanishes as $\hat{\zeta}\to\hat{\alpha}_j$.

For convenience we take the constants $\hat{\lambda}_s$ $(s=1,2,\cdots)$ to be the same set for each solution $\hat{W}_{2n+1,j}$ $(j=0,\pm 1)$, defining them recursively such that

$$\hat{a}_s=0 \qquad (s=1,2,\cdots). \tag{5.21}$$

The three solutions $\hat{W}_{2n+1,0}$, $\hat{W}_{2n+1,1}$, and $\hat{W}_{2n+1,-1}$ have the characteristic property of being recessive in the sectors $|\arg\hat{\zeta}|<\pi/3$, $\pi/3<\arg\hat{\zeta}<\pi$, and $-\pi<\arg\hat{\zeta}<-\pi/3$, respectively. Thus, we can identify them directly with standard Whittaker functions.

Beginning with $\hat{W}_{2n+1,0}(\kappa,\hat{\zeta})$, we identify this with the Whittaker function $(d\hat{\zeta}/d\xi)^{1/2}W_{\kappa,\mu}(\kappa\xi)$, since this too is a solution of (4.3) that is recessive in the sector $|\arg\hat{\zeta}|<\pi/3$. We have then the following identification:

$$W_{\kappa,\mu}(\kappa\xi)=\hat{c}_{2n+1,0}\left[\frac{(2\xi)^{1/2}\hat{\zeta}^{1/4}}{[(\xi-\xi_1)(\xi-\xi_2)]^{1/4}}\right]\hat{W}_{2n+1,0}(\kappa,\hat{\zeta}), \tag{5.22}$$

where $\hat{c}_{2n+1,0}$ is a constant of proportionality. We can determine the constant by comparing both sides of (5.22) as $\xi\to+\infty$ $(\hat{\zeta}\to\hat{\alpha}_0)$. On employing (4.7), (5.16), (5.20), and (5.21), as well as the asymptotic forms for $\mathrm{Ai}(z)$ and $\mathrm{Ai}'(z)$ with large argument (see [5, p. 392]), we find that the right-hand side of (5.22) is asymptotically equal to

$$\hat{c}_{2n+1,0}(2\pi)^{-1/2}\kappa^{-1/6}(2\xi)^{\kappa}\,e^{-\kappa\xi/2}\left(\frac{\xi_2(2-\alpha)}{2\xi_2-\alpha^2}\right)^{\kappa\alpha/2}(\xi_2-2)^{-\kappa}\left\{1-\sum_{s=0}^{n-1}\frac{\hat{b}_s}{\kappa^{2s+1}}\right\}. \tag{5.23}$$

This must be equal to the asymptotic form of $W_{\kappa,\mu}(\kappa\xi)$ as $\xi\to\infty$ (see (2.6)), and so we conclude that

$$\hat{c}_{2n+1,0}=(2\pi)^{1/2}\kappa^{1/6}\left(\frac{2\xi_2-\alpha^2}{\xi_2(2-\alpha)}\right)^{\kappa\alpha/2}\left(\kappa\left(\frac{\xi_2}{2}-1\right)\right)^{\kappa}\left\{1-\sum_{s=0}^{n-1}\frac{\hat{b}_s}{\kappa^{2s+1}}\right\}^{-1}. \tag{5.24}$$

Equation (5.22), with $\hat{W}_{2n+1,0}$ and $\hat{c}_{2n+1,0}$ given by (5.16) and (5.24), respectively, gives an asymptotic expansion for $W_{\kappa,\mu}(\kappa\xi)$ that is uniformly valid for $\hat{\zeta}\in\hat{\underline{\Delta}}^{(0)}$ and $\alpha\in[0,2-\delta]$, where $\hat{\underline{\Delta}}^{(0)}$ is the domain consisting of all points in $\hat{\underline{\Delta}}$ except those lying on either side of the cut interval BCB' (see Fig. 3(b)).

An asymptotic expansion for $W_{\kappa,\mu}(\kappa\xi)$ that is uniformly valid in the complementary ξ domain Δ can be obtained by replacing the functions $W_{-\kappa,\mu}(\kappa\xi\, e^{\pm\pi i})$ by their asymptotic forms (5.11) and (5.14) in the connection formula (2.9). For the result (stated in terms of the original parameters), see (6.8) in § 6.

The identification of the other two solutions in (5.16) is similarly achieved. For $j=1$ we have the identity

$$W_{-\kappa,\mu}(\kappa\xi\, e^{-\pi i})=\hat{c}_{2n+1,1}\left[\frac{(2\xi)^{1/2}\hat{\zeta}^{1/4}}{[(\xi-\xi_1)(\xi-\xi_2)]^{1/4}}\right]\hat{W}_{2n+1,1}(\kappa,\hat{\zeta}), \tag{5.25}$$

since both sides share the same recessive property in the sector $\pi/3<\arg\hat{\zeta}\leqq\pi$. By comparing both sides as $\zeta\to\hat{\alpha}_1$, $\xi\to-\infty+i0$ (see (4.7)), we obtain the following expression for the constant of proportionality:

$$\begin{aligned}\hat{c}_{2n+1,1}=(2\pi)^{1/2}\kappa^{1/6}\,e^{(\kappa-1/6)\pi i}&\left(\frac{\xi_2(2-\alpha)}{2\xi_2-\alpha^2}\right)^{\kappa\alpha/2}\\ &\cdot\left(\kappa\left(\frac{\xi_2}{2}-1\right)\right)^{-\kappa}\left\{1-e^{-\pi i/3}\sum_{s=0}^{n-1}\frac{\hat{b}_s}{\kappa^{2s+1}}\right\}^{-1}.\end{aligned} \tag{5.26}$$

This asymptotic expansion is uniformly valid for $\hat{\zeta} \in \hat{\underline{\Delta}}^{(1)}$ and $\alpha \in [0, 2-\delta]$, where $\hat{\underline{\Delta}}^{(1)}$ is the domain consisting of all points in $\hat{\underline{\Delta}}$ except those lying on the lower part of the cut interval BCB'.

Likewise, we can show that the following asymptotic expansion:

$$W_{-\kappa,\mu}(\kappa\xi e^{\pi i}) = \hat{c}_{2n+1,-1}\left[\frac{(2\xi)^{1/2}\hat{\zeta}^{1/4}}{[(\xi-\xi_1)(\xi-\xi_2)]^{1/4}}\right]\hat{W}_{2n+1,-1}(\kappa, \hat{\zeta}), \tag{5.27}$$

where

$$\hat{c}_{2n+1,-1} = (2\pi)^{1/2}\kappa^{1/6}\, e^{-(\kappa-1/6)\pi i}\left(\frac{\xi_2(2-\alpha)}{2\xi_2-\alpha^2}\right)^{\kappa\alpha/2} \cdot\left(\kappa\left(\frac{\xi_2}{2}-1\right)\right)^{-\kappa}\left\{1-e^{\pi i/3}\sum_{s=0}^{n-1}\frac{\hat{b}_s}{\kappa^{2s+1}}\right\}^{-1}, \tag{5.28}$$

is uniformly valid for $\hat{\zeta} \in \hat{\underline{\Delta}}^{(-1)}$ and $\alpha \in [0, 2-\delta]$, where $\hat{\underline{\Delta}}^{(-1)}$ is the domain consisting of all points in $\hat{\underline{\Delta}}$ except those lying on the upper part of the cut interval BCB'.

From (2.10), (5.25), and (5.27) we can obtain a uniform asymptotic expansion for $M_{\kappa,\mu}(\kappa\xi)$, uniformly valid for $\hat{\zeta} \in \hat{\underline{\Delta}}^{(0)}$, $\alpha \in [0, 2-\delta]$. The result, stated in terms of the original variables, is given in § 6, eq. (6.16).

6. Summary. For reference we now present the principal results of this paper, and in doing so we express them in terms of the original parameters κ and μ. We present asymptotic expansions as $\kappa \to \infty$ for the Whittaker functions $W_{\kappa,\mu}(z)$ and $M_{\kappa,\mu}(z)$. These expansions are uniformly valid for all real values of κ and μ such that

$$0 \leqq \mu/\kappa \leqq 1-\delta, \qquad \kappa > 0, \tag{6.1}$$

where δ is an arbitrarily small positive constant. Moreover the expansions, which either involve Airy functions or Bessel functions, are uniformly valid in certain complex z domains. These domains overlap and taken together cover the entire complex z plane. Thus an asymptotic expansion is available for both of the Whittaker functions for any value of z such that $|\arg z| \leqq \pi$. For other ranges of $\arg z$ we can use appropriate connection formulae (see, for example, Olver [5, p. 262]).

We also present expansions for the solutions of Whittaker's equation $W_{-\kappa,\mu}(z e^{-\pi i})$ and $W_{-\kappa,\mu}(z e^{\pi i})$. A separate identification for these functions is necessary because they are recessive in the second and third quadrants respectively of the complex z plane. Asymptotic expansions for the Whittaker function $M_{\kappa,\mu}(z)$ for negative values of κ are easily obtained from the ensuing results (where κ is positive) and the connection formula (2.11). Also, we can extend the following expansions to negative values of μ on using (2.2) and (2.12).

We first introduce two transformed variables ζ and $\hat{\zeta}$ that are related to the parameters κ, μ, and the independent variable z, with $-\pi < \arg z \leqq \pi$, by the following equations:

$$(4\mu^2-\kappa\zeta)^{1/2} - \mu \ln\left\{\frac{2\mu+(4\mu^2-\kappa\zeta)^{1/2}}{2\mu-(4\mu^2-\kappa\zeta)^{1/2}}\right\} = -\frac{1}{2}Z + \mu\ln\left\{\frac{z(\mu^2-\kappa^2+\kappa(\kappa^2-\mu^2)^{1/2})}{(\kappa-(\kappa^2-\mu^2)^{1/2})(2\mu^2-\kappa z-\mu Z)}\right\} + \kappa\ln\left\{\frac{2(\kappa^2-\mu^2)^{1/2}}{2\kappa-z+Z}\right\}, \tag{6.2}$$

$$\frac{2}{3}\kappa\hat{\zeta}^{3/2} = \frac{1}{2}Z + \mu\ln\left\{\frac{z(\kappa^2-\mu^2+\kappa(\kappa^2-\mu^2)^{1/2})}{(\kappa+(\kappa^2-\mu^2)^{1/2})(\kappa z-2\mu^2-\mu Z)}\right\} + \kappa\ln\left\{\frac{2(\kappa^2-\mu^2)^{1/2}}{z-2\kappa+Z}\right\}, \tag{6.3}$$

where

$$Z = (z^2-4\kappa z+4\mu^2)^{1/2}. \tag{6.4}$$

The fractional powers on the left side of (6.2) and (6.3) are assumed to take their principal values, with

$$-\pi<\arg\zeta\leqq\pi, \qquad -\pi<\arg\hat{\zeta}\leqq\pi.$$

Let $z=z_1$, z_2 denote the zeros $2\kappa-2(\kappa^2-\mu^2)^{1/2}$, $2\kappa+2(\kappa^2-\mu^2)^{1/2}$, respectively, of the function Z; we take the root of (6.4) such that Z is positive imaginary above the interval (z_1, z_2), negative imaginary below the same interval, and continuous elsewhere in the complex z plane.

For convenience we introduce the following functions:

$$\Phi(\kappa,\mu,z)=\left(\frac{4\mu^2-\kappa^2\zeta}{z_1-z}\right)^{1/4}(z_2-z)^{-1/4}\left(\frac{z}{\kappa}\right)^{1/2}, \tag{6.5}$$

$$\hat{\Phi}(\kappa,\mu,z)=\left(\frac{\hat{\zeta}}{z-z_2}\right)^{1/4}\frac{(2z)^{1/2}}{(z-z_1)^{1/4}}, \tag{6.6}$$

where again principal values are to be taken.

The following asymptotic expansions are uniformly valid for $0\leqq\mu/\kappa\leqq 1-\delta$, $\kappa>0$, and for z and ζ lying in the specified complex domains.

$$M_{\kappa,\mu}(z)=\kappa^{1/2}\Gamma(2\mu+1)\left(\frac{e^2}{\kappa^2-\mu^2}\right)^{\mu/2}\left(\frac{\kappa-\mu}{\kappa+\mu}\right)^{\kappa/2}\left\{1+2\mu\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{\kappa^{2s+2}}\right\}^{-1}$$

$$\cdot\Phi(\kappa,\mu,z)\left[J_{2\mu}(\kappa\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\zeta)}{\kappa^{2s}}+\frac{\zeta^{1/2}}{\kappa}J'_{2\mu}(\kappa\zeta^{1/2})\right. \tag{6.7}$$

$$\left.\cdot\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\kappa^{2s}}+\frac{1}{2}\zeta^{-1/2}\varepsilon^{(0)}_{2n+1}(\zeta)\right], \qquad \frac{z}{\kappa}\in\Delta^{(0)}, \quad \zeta\in\underline{\Delta}^{(0)};$$

$$W_{\kappa,\mu}(z)=\frac{1}{\pi}\Gamma\left(\kappa+\mu+\frac{1}{2}\right)\Gamma\left(\kappa-\mu+\frac{1}{2}\right)e^{-\mu\pi i}c^{(1)}_{2n+1}\Phi(\kappa,\mu,z)$$

$$\cdot\left[\{J_{2\mu}(\kappa\zeta^{1/2})\sin(\kappa-\mu)\pi-Y_{2\mu}(\kappa\zeta^{1/2})\cos(\kappa-\mu)\pi\}\sum_{s=0}^{n}\frac{A_s(\zeta)}{\kappa^{2s}}\right.$$

$$+\frac{\zeta^{1/2}}{\kappa}\{J'_{2\mu}(\kappa\zeta^{1/2})\sin(\kappa-\mu)\pi-Y'_{2\mu}(\kappa\zeta^{1/2})\cos(\kappa-\mu)\pi\}\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\kappa^{2s}} \tag{6.8}$$

$$\left.+\frac{i}{2}\zeta^{-1/2}\{e^{-(\kappa-\mu)\pi i}\varepsilon^{(1)}_{2n+1}(\zeta)-e^{(\kappa-\mu)\pi i}\varepsilon^{(2)}_{2n+1}(\zeta)\}\right],$$

$$\frac{z}{\kappa}\in\Delta, \quad \zeta\in\underline{\Delta},$$

$$W_{-\kappa,\mu}(z\,e^{-\pi i})=c^{(1)}_{2n+1}\Phi(\kappa,\mu,z)\left[H^{(1)}_{2\mu}(\kappa\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\zeta)}{\kappa^{2s}}\right.$$

$$\left.+\frac{\zeta^{1/2}}{\kappa}H^{(1)'}_{2\mu}(\kappa\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\kappa^{2s}}+\zeta^{-1/2}\varepsilon^{(1)}_{2n+1}(\zeta)\right], \tag{6.9}$$

$$\frac{z}{\kappa}\in\Delta, \quad \zeta\in\underline{\Delta},$$

$$W_{-\kappa,\mu}(z\,e^{\pi i}) = c_{2n+1}^{(1)}\, e^{-2\mu\pi i}\Phi(\kappa,\mu,z)\Bigg[H_{2\mu}^{(2)}(\kappa\zeta^{1/2}) \sum_{s=0}^{n} \frac{A_s(\zeta)}{\kappa^{2s}}$$

$$+\frac{\zeta}{\kappa} H_{2\mu}^{(2)'}(\kappa\zeta^{1/2}) \sum_{s=0}^{n-1} \frac{B_s(\zeta)}{\kappa^{2s}} + \zeta^{-1/2}\varepsilon_{2n+1}^{(2)}(\zeta)\Bigg], \tag{6.10}$$

$$\frac{z}{\kappa}\in\Delta, \quad \zeta\in\underline{\Delta}.$$

The coefficient $c_{2n+1}^{(1)}$ is given by

$$c_{2n+1}^{(1)} = -e^{\mu\pi i}\left(\frac{\pi\kappa}{2}\right)^{1/2}\left(\frac{\kappa-\mu}{\kappa+\mu}\right)^{\mu/2}(\kappa^2-\mu^2)^{-\kappa/2}\left\{1-\sum_{s=0}^{n-1}\frac{b_s}{\kappa^{2s+1}}\right\}^{-1}. \tag{6.11}$$

The z domain of validity Δ, and the corresponding ζ domain $\underline{\Delta}$, are illustrated in Figs 1(a) and 1(b). The domain $\Delta^{(0)}$ ($\underline{\Delta}^{(0)}$) consists of all points in Δ ($\underline{\Delta}$) except those lying on the lines CD, CD'. The coefficients $A_s(\zeta)$ and $B_s(\zeta)$ are defined recursively by (5.2) and (5.3a, b), with (5.6) applying to (6.7), and (5.5) applying to (6.8), (6.9), and (6.10). In (5.2) the function ψ is given by (3.5) together with (1.3) and (1.4).

The error terms $\varepsilon_{2n+1}^{(0)}$, $\varepsilon_{2n+1}^{(1)}$, and $\varepsilon_{2n+1}^{(2)}$ have explicit upper bounds available (see [2, eq. (5.16)]). From these bounds the following asymptotic properties as $\kappa\to\infty$ can be deduced:

$$\zeta^{-1/2}\varepsilon_{2n+1}^{(j)}(\zeta) = \mathscr{C}_{2\mu}^{(j)}(\kappa\zeta^{1/2})O(\kappa^{-2n-1}) + \frac{\zeta^{1/2}}{\kappa}\mathscr{C}_{2\mu}^{(j)'}(\kappa\zeta^{1/2})O(\kappa^{-2n+1}) \quad (j=0,1,2), \tag{6.12}$$

where $\mathscr{C}_\nu^{(j)} = 2J_\nu$, $H_\nu^{(1)}$, $H_\nu^{(2)}$ for $j=0,1,2$, respectively. The O-terms are uniform for $0\leqq\mu/\kappa\leqq 1-\delta$, $\kappa>0$, and ζ lying in the respective domain of validity.

The error terms also satisfy the following boundary conditions:

$$\lim_{\zeta\to 0}\{\varepsilon_{2n+1}^{(0)}(\zeta)/J_{2\mu}(\kappa\zeta^{1/2})\} = 0, \tag{6.13}$$

$$\lim_{\zeta\to -\infty+i0}\{\varepsilon_{2n+1}^{(1)}(\zeta)/H_{2\mu}^{(1)}(\kappa\zeta^{1/2})\} = 0, \tag{6.14}$$

$$\lim_{\zeta\to -\infty-i0}\{\varepsilon_{2n+1}^{(2)}(\zeta)/H_{2\mu}^{(2)}(\kappa\zeta^{1/2})\} = 0. \tag{6.15}$$

The following asymptotic expansions are uniformly valid for $0\leqq\mu/\kappa\leqq 1-\delta$, $\kappa>0$, and for z and $\hat{\zeta}$ lying in the specified complex domains:

$$M_{\kappa,\mu}(z) = \frac{\hat{c}}{2\pi}\Gamma(2\mu+1)\Gamma\left(\kappa-\mu+\frac{1}{2}\right)\left(1+\hat{\phi}_n^2-\frac{1}{2}\hat{\phi}_n\right)^{-1}\hat{\Phi}(\kappa,\mu,z)$$

$$\cdot\Bigg[\left\{Ai(\kappa^{2/3}\hat{\zeta})\left(\sin(\kappa-\mu)\pi-\hat{\phi}_n\sin\left(\kappa-\mu+\frac{1}{3}\right)\pi\right)\right.$$

$$\left.+Bi(\kappa^{2/3}\hat{\zeta})\left(\cos(\kappa-\mu)\pi-\hat{\phi}_n\cos\left(\kappa-\mu+\frac{1}{3}\right)\pi\right)\right\}\sum_{s=0}^{n}\frac{\hat{A}_s(\hat{\zeta})}{\kappa^{2s}}$$

$$-\kappa^{-4/3}\left\{Ai'(\kappa^{2/3}\hat{\zeta})\left(\sin\left(\kappa-\mu-\frac{1}{3}\right)\pi-\hat{\phi}_n\sin(\kappa-\mu)\pi\right)\right. \tag{6.16}$$

$$\left.+Bi'(\kappa^{2/3}\hat{\zeta})\left(\cos\left(\kappa-\mu-\frac{1}{3}\right)\pi-\hat{\phi}_n\cos(\kappa-\mu)\pi\right)\right\}\sum_{s=0}^{n-1}\frac{\hat{B}_s(\hat{\zeta})}{\kappa^{2s}}$$

$$+\{e^{(\kappa-\mu-1/6)\pi i}(1-e^{\pi i/3}\hat{\phi}_n)\hat{\varepsilon}_{2n+1,1}(\hat{\zeta}) + e^{-(\kappa-\mu-1/6)\pi i}(1-e^{-\pi i/3}\hat{\phi}_n)\hat{\varepsilon}_{2n+1,-1}(\hat{\zeta})\}\Big],$$

$$\frac{z}{\kappa}\in\hat{\Delta}^{(0)}, \quad \hat{\zeta}\in\hat{\underline{\Delta}}^{(0)},$$

$$W_{\kappa,\mu}(z)=(2\pi)^{1/2}\kappa^{1/6}\left(\frac{\kappa+\mu}{\kappa-\mu}\right)^{\mu/2}(4(\kappa^2-\mu^2))^{\kappa/2}(1-\hat{\phi}_n)^{-1}\hat{\Phi}(\kappa,\mu,z)$$

$$\cdot\Big[\mathrm{Ai}\,(\kappa^{2/3}\hat{\zeta})\sum_{s=0}^{n}\frac{\hat{A}_s(\hat{\zeta})}{\kappa^{2s}}+\frac{\mathrm{Ai}'\,(\kappa^{2/3}\hat{\zeta})}{\kappa^{4/3}}\sum_{s=0}^{n-1}\frac{\hat{B}_s(\hat{\zeta})}{\kappa^{2s}}+\hat{\varepsilon}_{2n+1,0}(\hat{\zeta})\Big], \tag{6.17}$$

$$\frac{z}{\kappa}\in\hat{\Delta}^{(0)}, \quad \hat{\zeta}\in\hat{\underline{\Delta}}^{(0)};$$

$$W_{-\kappa,\mu}(z\,e^{-\pi i})=\hat{c}\,e^{(\kappa-1/6)\pi i}(1-e^{-\pi i/3}\hat{\phi}_n)^{-1}\hat{\Phi}(\kappa,\mu,z)$$

$$\cdot\Big[\mathrm{Ai}\,(\kappa^{2/3}\hat{\zeta}\,e^{-2\pi i/3})\sum_{s=0}^{n}\frac{\hat{A}_s(\hat{\zeta})}{\kappa^{2s}} +\frac{\mathrm{Ai}'\,(\kappa^{2/3}\hat{\zeta}\,e^{-2\pi i/3})}{\kappa^{4/3}}\sum_{s=0}^{n-1}\frac{\hat{B}_s(\hat{\zeta})}{\kappa^{2s}}+\hat{\varepsilon}_{2n+1,1}(\hat{\zeta})\Big], \tag{6.18}$$

$$\frac{z}{\kappa}\in\hat{\Delta}^{(1)}, \quad \hat{\zeta}\in\hat{\underline{\Delta}}^{(1)},$$

$$W_{-\kappa,\mu}(z\,e^{\pi i})=\hat{c}\,e^{-(\kappa-1/6)\pi i}(1-e^{\pi i/3}\hat{\phi}_n)^{-1}\hat{\Phi}(\kappa,\mu,z)$$

$$\cdot\Big[\mathrm{Ai}\,(\kappa^{2/3}\hat{\zeta}\,e^{2\pi i/3})\sum_{s=0}^{n}\frac{\hat{A}_s(\hat{\zeta})}{\kappa^{2s}} +\frac{\mathrm{Ai}'\,(\kappa^{2/3}\hat{\zeta}\,e^{2\pi i/3})}{\kappa^{4/3}}\sum_{s=0}^{n-1}\frac{\hat{B}_s(\hat{\zeta})}{\kappa^{2s}}+\hat{\varepsilon}_{2n+1,-1}(\hat{\zeta})\Big] \tag{6.19}$$

$$\frac{z}{\kappa}\in\hat{\Delta}^{(-1)}, \quad \hat{\zeta}\in\hat{\underline{\Delta}}^{(-1)}.$$

The coefficient $\hat{c}$ is given by

$$\hat{c}=(2\pi)^{1/2}\kappa^{1/6}\left(\frac{\kappa-\mu}{\kappa+\mu}\right)^{\mu/2}(4(\kappa^2-\mu^2))^{-\kappa/2}, \tag{6.20}$$

and $\hat{\phi}_n$ is defined by

$$\hat{\phi}_n=\sum_{s=0}^{n-1}\frac{\hat{b}_s}{\kappa^{2s+1}}. \tag{6.21}$$

The z domain $\hat{\Delta}$ and the corresponding $\hat{\zeta}$ domain $\hat{\underline{\Delta}}$ are illustrated in Figs. 3(a) and 3(b). The $\hat{\zeta}$ domains of validity $\hat{\underline{\Delta}}^{(1)}$ and $\hat{\underline{\Delta}}^{(-1)}$ consist of all points in $\hat{\underline{\Delta}}$ except those lying, respectively, on the lower and upper parts of the cut interval BCB'. $\hat{\underline{\Delta}}^{(0)}$ is the intersection of $\hat{\underline{\Delta}}^{(1)}$ and $\hat{\underline{\Delta}}^{(-1)}$.

The coefficients $\hat{A}_s(\hat{\zeta})$ and $\hat{B}_s(\hat{\zeta})$ are defined recursively by (5.17), (5.18), with (4.4) and (5.21). Bounds for the error functions $\hat{\varepsilon}_{2n+1}$, $\hat{\varepsilon}_{2n+1,1}$, $\hat{\varepsilon}_{2n+1,-1}$, are supplied in [5, p. 418]. These functions satisfy the following boundary conditions:

$$\lim_{\hat{\zeta}\to\infty e^{2\pi ij/3}}\{\hat{\varepsilon}_{2n+1,j}(\hat{\zeta})/\mathrm{Ai}\,(\kappa^{2/3}\hat{\zeta}\,e^{-2\pi ij/3})\}=0 \qquad (j=0,\pm1). \tag{6.22}$$

Finally, we remark that (6.8) and (6.16) should not be used when $\kappa-\mu-\frac{1}{2}$ is equal to a nonnegative integer. In this case the asymptotic formulae (6.7) and (6.17) should be used, together with the relation

$$W_{\kappa,\mu}(z)=\frac{\Gamma(-2\mu)}{\Gamma(\frac{1}{2}-\mu-\kappa)}M_{\kappa,\mu}(z). \tag{6.23}$$

REFERENCES

[1] G. B. BAUMGARTNER, *Uniform asymptotic approximations for the Whittaker function $M_{k,m}(z)$*, Ph.D. thesis, Illinois Institute of Technology, Chicago, IL, 1980.

[2] W. G. C. BOYD AND T. M. DUNSTER, *Uniform asymptotic solutions of a class of second-order linear differential equations having a turning point and a regular singularity, with an application to Legendre functions*, SIAM J. Math. Anal., 17 (1986), pp. 422-450.

[3] A. ERDÉLYI AND C. A. SWANSON, *Asymptotic Forms of Whittaker's Confluent Hypergeometric Functions*, Mem. Amer. Math. Soc., 25, New York, 1957.

[4] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series and Products*, Fourth Edition, Academic Press, London, 1980.

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

[6] ———, *Unsolved problems in the asymptotic estimation of special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 99-141.

[7] ———, *Second-order linear differential equations with two turning points*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 137-174.

[8] ———, *Whittaker functions with both parameters large: Uniform approximations in terms of parabolic cylinder functions*, Proc. Roy. Soc. Edinburgh Sec. A, (1980), pp. 213-234.

[9] H. SKOVGAARD, *Uniform Asymptotic Expansions of Confluent Hypergeometric Functions and Whittaker Functions*, Gjellerups, Copenhagen, 1966.

SIAM J. MATH. ANAL.
Vol. 20, No. 3, pp. 761-762, May 1989

018

ERRATUM:

Study of a Doubly Nonlinear Heat Equation with No Growth Assumptions on the Parabolic Term*

D. BLANCHARD† AND G. A. FRANCFORT†

Inequalities (64) and (101) of [1] are not satisfied under the assumptions on Φ listed in (5) of that paper. Rather, they immediately result from the following implicitly (*but not explicitly*) assumed almost pointwise inequality:

$$(I) \qquad (D\Phi(w)(x) - D\Phi(w')(x),\, w(x) - w'(x))_{\mathbb{R}^N} \geqq 0,$$

for any w and w' in $[L_q(\Omega)]^N$.

Counterexamples of functionals Φ that satisfy (5) of [1] and for which (I) is violated are easily obtained. In the usually considered case where Φ is a C^1, convex, coercive, *local* functional (cf. definition below), (I) is trivially satisfied and inequalities (64), (101) hold true.

DEFINITION. The functional Φ defined on $[L_q(\Omega)]^N$, $q > 1$, is a C^1, convex, coercive, local functional if there exists a convex normal integrand $f(x, \xi)$ on $\Omega \times \mathbb{R}^N$, C^1 in ξ for almost every x in Ω with the following properties:

There exist an element $a(x)$ of $L_1(\Omega)$, two strictly positive constants α and β such that, for almost every x of Ω and every ξ of $\mathbb{R}^N$,

$$\alpha|\xi|^q \leqq f(x, \xi) \leqq a(x) + \beta|\xi|^q,$$

For every w in $[L_q(\Omega)]^N$,

$$\Phi(w) = \int_\Omega f(x, w(x))\, dx.$$

Remark. If Φ is a C^1, convex, coercive, local functional and if $\Phi(0) = 0$, $D\Phi$ is bounded on the bounded sets of $[L_q(\Omega)]^N$ and Φ satisfies (5) of [1] with a coercivity exponent r equal to q.

To ensure the validity of the results of our paper, it is then tempting to impose (I) as an additional hypothesis to (5) and to investigate the nonlocal functionals that satisfy (5) and (I). In fact there are *no* such functionals and the following theorem can be deduced from the work of Buttazzo and Dal Maso concerning integral representations of local functionals (the proof uses [2, Thm. 1.4 and Cors. 1.5, 1.7]).

THEOREM. *If Φ satisfies* (5) *and* (I), *Φ is a C^1, convex, coercive, local functional on $[L_q(\Omega)]^N$.*

Conclusion. If Φ is a C^1, convex, coercive, local functional on $[L_q(\Omega)]^N$ in the sense of the above definition (and if $\Phi(0) = 0$), Theorems 1 and 2 of [1] hold true as stated. In fact, Theorems 1 and 2 hold true (*except for the comparison results*) even when (I) is not satisfied, as can be shown by the method developed for the proof of Theorem 2, but such considerations are beyond the scope of an erratum.

* Received by the editors October 8, 1988; accepted October 21, 1988.
† Laboratoire Central des Ponts et Chausées, 58 boulevard Lefebvre, 75732 Paris Cedex 15, France

REFERENCES

[1] D. BLANCHARD AND G. A. FRANCFORT, *Study of a doubly nonlinear heat equation with no growth assumption on the parabolic term*, SIAM J. Math. Anal., 19 (1988), pp. 1032–1056.

[2] G. BUTTAZZO AND G. DAL MASO, *On Nemyckii operators and integral representation of local functionals*, Rend. Mat., 7 (1983), pp. 491–509.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 763-781, July 1989

REGULARITY OF THE SOLUTION OF ELLIPTIC PROBLEMS WITH PIECEWISE ANALYTIC DATA, II: THE TRACE SPACES AND APPLICATION TO THE BOUNDARY VALUE PROBLEMS WITH NONHOMOGENEOUS BOUNDARY CONDITIONS*

I. BABUŠKA† AND B. Q. GUO‡

Abstract. This paper analyzes the trace spaces of the weighted space $\mathfrak{B}^2_\beta(\Omega)$ introduced by Babuška and Guo [*SIAM J. Math. Anal.*, 19 (1988), pp. 172-203].

Key words. elliptic equation with piecewise analytic data, Dirichlet problem, corner singularities

AMS(MOS) subject classifications. 35B65, 35D10, 35G15, 35J05

1. Introduction. Elliptic boundary value problems with piecewise analytic data are typical in many fields of applications, for example, in structural mechanics. These problems are then numerically analyzed in engineering by the finite-element method. The design and performance of a numerical method directly depends on the class of problems to which it is oriented. The smaller the class is, the more effective the numerical method can be. Hence, it is important to characterize mathematically a (minimal) class that encompasses virtually all practical problems in a field of applications. The space $\mathfrak{B}^2_\beta(\Omega)$ is such a class. In [4], [5a], and [5b] it has been shown that if the solution belongs to the space $\mathfrak{B}^2_\beta(\Omega)$, then the *h-p* version of the finite-element method has an exponential rate of convergence. The *h-p* version uses properly refined mesh and a high degree of elements in contrast to the usual *h*-version that uses only low-degree elements. For the survey of various theoretical and practical aspects of the *h-p* version we refer the reader to [1] and the references given therein.

In [3] the spaces $\mathfrak{B}^2_\beta(\Omega)$ have been analyzed. It has been shown that the solution of the elliptic boundary value problems with piecewise analytic data belongs to these spaces.

The present paper elaborates in detail on the structure of the traces of functions of $\mathfrak{B}^2_\beta(\Omega)$. The results give easy characterization of the case when the solution belongs to $\mathfrak{B}^2_\beta(\Omega)$. In § 2 we give the preliminaries and basic definitions. Section 3 defines the model problem of second-order elliptic partial differential equations. Section 4 introduces the space of traces of $u \in \mathfrak{B}^2_\beta(\Omega)$ on the boundary $\partial\Omega$. It is also shown that these traces can be extended into $\mathfrak{B}^2_\beta(\Omega)$.

2. Preliminaries. Let $\Omega \subset \mathbb{R}^2$, $(x_1, x_2) = x$ be a simply-connected, bounded domain with the boundary $\partial\Omega = \Gamma = \cup_{i=1}^M \bar{\Gamma}_i$. Γ_i are analytic simple arcs called *edges*,

$$\bar{\Gamma}_i \in \{(\varphi_i(\xi), \psi_i(\xi)) \,|\, \xi \in \bar{I} = [-1, 1]\},$$

where $\varphi_i(\xi)$, $\psi_i(\xi)$ are *analytic functions* on $\bar{I}$ and $|\varphi_i'(\xi)|^2 + |\psi_i'(\xi)|^2 \geqq \alpha_i > 0$. By Γ_i we denote the open arc, i.e., the image of $I = (-1, 1)$. Let A_i, $i = 1, \cdots, M$, be the vertices

* Received by the editors April 6, 1988; accepted for publication October 11, 1988.

† Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The work of this author was supported by the Office of Naval Research under contract N00014-85-K-0169.

‡ Engineering Mechanics Research Corporation, Troy, Michigan. The work of this author was supported by the National Science Foundation under grant DMS-85-16191 during a stay at the Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

of Ω and $\Gamma_i = A_iA_{i+1}$, i.e., the edge Γ_i is linking the vertices A_i and A_{i+1}. For simplicity we will also write $A_1 = A_{M+1}$. An example of the domain Ω under consideration is given in Fig. 2.1. By ω_i, $i = 1, \cdots, M$, we denote the internal angles of Ω at A_i. We will assume that $0 < \omega_i \leqq 2\pi$. We will also consider the case when two edges coincide. Then we understand them in a "two-sided" sense. If all edges are straight lines then we call the domain Ω a *straight polygon.* Otherwise we speak about a *curvilinear polygon.* If $0 < \omega_i < 2\pi$, $i = 1, \cdots, M$, we speak about a Lipschitzian domain. Let us assume that $\Gamma = \Gamma^{(0)} \cup \Gamma^{(1)}$ where $\Gamma^{(0)} = \cup_{i \in Q} \bar{\Gamma}_i$, $\Gamma^{(1)} = \Gamma - \Gamma^{(0)}$, $\bar{\Gamma}^{(1)} = \cup_{i \in Q'} \bar{\Gamma}_{i'}$, where Q is some subset of the set $\{1, 2, \cdots, M\} = \mathcal{M}$ and $Q' = \mathcal{M} - Q$.

We assume for simplicity that Ω is a simply-connected domain. The results we present here are also valid when Ω is an n-connected, bounded domain and its boundary is composed of n-curves.

Denote $I = \{x \mid -1 < x < 1\}$; we also write $I = \{x_1, x_2 \mid -1 < x_1 < 1, x_2 = 0\} \subset \mathbb{R}_2$ when no misunderstanding can occur.

By $L_2(\Omega)$, $L_p(\Omega)$, $L_2(I)$, $L_p(I)$, the usual spaces of p-integrable, $1 < p < \infty$, functions on Ω or I are denoted. By $H^m(\Omega)$, $H^m(I)$, $m \geqq 0$ integer, we denote the usual Sobolev space of functions with square integrable derivatives of order less than or equal to m on Ω (respectively, I). The space $H^m(\Omega)$ is furnished with the usual norm

$$\|u\|^2_{H^m(\Omega)} = \sum_{0 \leqq |\alpha| \leqq m} \|D^\alpha u\|^2_{L_2(\Omega)}$$

where $\alpha = (\alpha_1, \alpha_2)$, $\alpha_i \geqq 0$ integer, $i = 1, 2$, $|\alpha| = \alpha_1 + \alpha_2$, and

$$D^\alpha u = \frac{\partial^\alpha u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} = u_{x_1^{\alpha_1} x_2^{\alpha_2}}.$$

Furthermore, we let

$$|u|_{H^m(\Omega)} = \| |D^m u| \|_{L_2(\Omega)}, \qquad |D^m u|^2 = \sum_{|\alpha| = m} |D^\alpha u|^2.$$

As usual, we write $H^0(\Omega) = L_2(\Omega)$,

$$H^1_0(\Omega) = \{u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma^{(0)}\}.$$

In an analogous way we define $H^m(I)$ by $D^k u = u^{(k)} = d^k u / dx^k$.

By $r_j(x) = \text{dist}\,(x, A_j) = |x - A_j|$, $x \in \Omega$, $j \in \mathcal{M}$, we denote the Euclidean distance between the point x and the vertex A_j, $\hat{r}_1(x) = x + 1$, $\hat{r}_2(x) = x - 1$, $x \in I$. Let $\beta = (\beta_1, \cdots, \beta_M)$ (respectively, $\beta = (\beta_1, \beta_2)$) be an M-tuple of real numbers $0 < \beta_i < 1$,

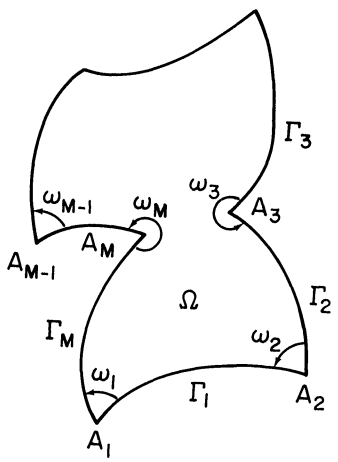


FIG. 2.1. *The scheme of the domain.*

$i=1,\cdots,M$. We write $\alpha_1<\beta<\alpha_2$ (respectively, $\bar{\beta}<\beta$) if $\alpha_1<\beta_i<\alpha_2$ (respectively, $\bar{\beta}_i<\beta_i$), $i=1,\cdots,M$. For any integer β, we write $\beta+k=\{\beta_1+k,\cdots,\beta_M+k\}$ such that

$$\Phi_{\beta+k}(x)=\prod_{i=1}^{M}|r_i(x)|^{\beta_i+k},\qquad x\in\Omega$$

and

$$\hat{\Phi}_{\beta+k}(x)=\prod_{i=1}^{2}|\hat{r}_i(x)|^{\beta_i+k},\qquad x\in I.$$

By $C^j(\Omega)$, $C^j(\bar{\Omega})$, $C^j(I)$, $C^j(\bar{I})$, $j\geqq 0$ integer we will denote the set of all functions with continuous j-derivatives on Ω, $\bar{\Omega}$, I, $\bar{I}$, furnished with the usual norm $\|\cdot\|_{C^j(\Omega)}$, $\|\cdot\|_{C^j(I)}$. Let $H_\beta^{m,l}(\Omega)$, $m\geqq l\geqq 0$ integers, be the completion of the set of all infinitely differentiable functions under the norm

$$\|u\|^2_{H_\beta^{m,l}(\Omega)}=\|u\|^2_{H^{l-1}(\Omega)}+\sum_{\substack{|\alpha|=k\\ k=l}}^{k=m}\|\Phi_{\beta+k-l}|D^\alpha u|\,\|^2_{L_2(\Omega)}\quad\text{for } l\geqq 1,$$

$$\|u\|^2_{H_\beta^{m,0}(\Omega)}=\sum_{\substack{|\alpha|=k\\ k=0}}^{k=m}\|\Phi_{\beta+k}|D^\alpha u|\,\|^2_{L_2(\Omega)}.$$

If $m=l=0$ we will write $H_\beta^{0,0}=L_\beta(\Omega)$. Analogously as before we define

$$|u|^2_{H_\beta^{l,l}(\Omega)}=\sum_{|\alpha|=l}\|\Phi_\beta|D^\alpha u|\,\|^2_{L_2(\Omega)}.$$

In a similar way $H_\beta^{m,l}(I)$ is defined

$$\|u\|^2_{H_\beta^{m,l}(I)}=\|u\|^2_{H^{l-1}(I)}+\sum_{k=l}^{k=m}\|\hat{\Phi}_{\beta+k-l}|D^k u|\,\|^2_{L_2(I)}\quad\text{for } l\geqq 1,$$

$$\|u\|^2_{H_\beta^{m,0}(I)}=\sum_{k=0}^{k=m}\|\hat{\Phi}_{\beta+k}|D^k u|\,\|^2_{L_2(I)}.$$

Furthermore we introduce the space $\mathfrak{B}_\beta^l(\Omega)$, $l\geqq 0$ integer that will play an important role in this paper:

$$\mathfrak{B}_\beta^l(\Omega)=\{u\,|\,u\in H_\beta^{k,l}(\Omega),\text{ for any } k\geqq l,\ \|\Phi_{\beta+k-l}|D^\alpha u|\,\|_{L_2(\Omega)}\leqq Cd^{k-l}(k-l)!,\ |\alpha|=k,\ C>0,\ d\geqq 1 \text{ independent of } k\},$$

where C and d may depend on u. If we wish to emphasize the dependence on d we will write $\mathfrak{B}_{\beta,d}^l(\Omega)$. Analogously for $l\geqq 0$ integer

$$\mathfrak{B}_\beta^l(I)=\{u\,|\,u\in H_\beta^{k,l}(I),\text{ for any } k\geqq l,\ \|\hat{\Phi}_{\beta+k-l}u^{(k)}\|_{L_2(I)}\leqq Cd^{k-l}(k-l)!,\ C>0,\ d\geqq 1 \text{ independent of } k\}.$$

Furthermore, for $j=1,2$,

$$\mathfrak{C}_\beta^j(\Omega)=\{u\in H_\beta^{j,j}(\Omega)\,|\,|D^\alpha u(x)|\leqq Cd^k k!|\Phi_{k+\beta-j+1}(x)|^{-1},\ |\alpha|=k=j-1,j,\cdots,\ C>0,\ d\geqq 1 \text{ independent of } k\},$$

$$\mathfrak{C}_\beta^j(I)=\{u\in H_\beta^{j,j}(I)\,|\,|u^{(k)}(x)|\leqq C|\hat{\Phi}_{k+\beta-j+1/2}(x)|^{-1}d^k k!,\ k=j-1,j,\cdots,\ C>0,\ d\geqq 1 \text{ independent of } k\}.$$

Let $\gamma = \cup_{i \in \rho \subset \mathcal{M}} \bar{\Gamma}_i$. Then we define $H^{k-1/2}(\gamma)$, $k \geqq 1$ (respectively, $H_\beta^{k-1/2,l-1/2}(\gamma)$, $k \geqq l \geqq 1$) integers as follows: for any $\varphi \in H^{k-1/2}(\gamma)$ (respectively, $H_\beta^{k-1/2,l-1/2}(\gamma)$) there exists $f \in H^k(\Omega)$ (respectively, $H_\beta^{k,l}(\Omega)$) such that $f|_\gamma = \varphi$. We define then

$$\|\varphi\|_{H^{k-1/2}(\Gamma)} \quad \text{(respectively, } \|\varphi\|_{H_\beta^{k-1/2,l-1/2}(\gamma)})$$
$$= \inf_{f|_\gamma = \varphi} \|f\|_{H^k(\Omega)} \quad \text{(respectively, } \|f\|_{H_\beta^{k,l}(\Omega)}).$$

By $\mathfrak{B}_\beta^{l-1/2}(\gamma)$, $l \geqq 1$, we will denote the set of the traces on γ of functions from the space $\mathfrak{B}_\beta^l(\Omega)$.

Let Γ_i be an edge of Ω; then by the assumption there exists a one-to-one mapping m_i of I onto Γ_i which is analytic. If Γ_i is a straight line, then we will assume that m_i is the linear mapping. Let u be defined on Γ_i, $U(x) = u(m_i(x))$ be defined on I. Then we define

$$H^m(\Gamma_i) = \{u \mid U \in H^m(I)\}, \qquad \|u\|_{H^m(\Gamma_i)} = \|U\|_{H^m(I)}.$$

In the same way we define the spaces $H_\beta^{m,l}(\Gamma_i)$, $\mathfrak{B}_\beta^l(\Gamma_i)$, $\mathfrak{C}_\beta^l(\Gamma_i)$. Let us remark that, as we defined it, $\|\cdot\|_{H^m(\Gamma_i)}$ depends on the mapping m_i, i.e., it depends on the parameterization of the arc Γ_i. Nevertheless the space $H_\beta^{m,l}(\Gamma_i)$ does not do as well as $\mathfrak{C}_\beta^l(\Gamma_i)$ (see Lemma 4.6) but $\mathfrak{B}_\beta^l(\Gamma_i)$ could be dependent on m_i. Let us now state some lemmas that will be used later.

LEMMA 2.1. *We have*

$$H_\beta^{2,2}(\Omega) \subset C^0(\bar{\Omega})$$

with the continuous injection.

See Lemma 7 of [2].

LEMMA 2.2. *Let* $u \in H_\beta^{2,2}(\Omega)$. *Then*

(i)

$$\| |D^1 u| \Phi_{\beta-1} \|_{L_2(\Omega)} \leqq C \|u\|_{H_\beta^{2,2}(\Omega)}. \tag{2.1}$$

(ii) *Let* $u(A_i) = 0$, $i = 1, \cdots, M$. *Then*

$$\|u \Phi_{\beta-2}\|_{L_2(\Omega)} \leqq C \|u\|_{H_\beta^{2,2}(\Omega)}. \tag{2.2}$$

See Lemma 8 of [2].

LEMMA 2.3. $\mathfrak{B}_\beta^2(\Omega) \subset \mathfrak{C}_\beta^2(\Omega)$ *and* $\mathfrak{C}_\beta^2(\Omega) \subset \mathfrak{B}_{\beta+\varepsilon}^2(\Omega)$, $0 < \beta + \varepsilon < 1$, $\varepsilon > 0$ *arbitrary.*

See Theorems 2.2 and 2.3 of [4].

LEMMA 2.4. *Let* $u \in \mathfrak{B}_\beta^j(\Omega)$, $j \geqq 0$; *then* u *is analytic on* $\bar{\Omega} - \cup_{i=1}^M A_i$.

LEMMA 2.5. *Let* $r \neq 1$ *and* $F(x)$, $0 < x < \infty$ *is defined by*

$$F(x) = \int_0^x f(t)\, dt \quad \text{for } r > 1,$$

$$F(x) = \int_x^\infty f(t)\, dt \quad \text{for } r < 1.$$

Then

$$\int_0^\infty x^{-r} F^2(x)\, dx \leqq \left(\frac{2}{|r-1|}\right)^2 \int_0^\infty x^{-r} (xf)^2\, dx.$$

See Theorem 330 of [7].

3. The model problem and its properties. Let Ω be the curvilinear or straight polygon and let L be a strongly elliptic operator

$$L(u) = -\sum_{i,j=1}^{2} (a_{i,j}(x)u_{x_i})_{x_j} + \sum_{i=1}^{2} b_i(x)u_{x_i} + c(x)u$$

where $a_{i,j}(x) = a_{j,i}(x)$, $b_i(x)$, $c(x)$ are analytic functions on $\bar{\Omega}$ and for any $\xi_1, \xi_2 \in \mathbb{R}$ and any $x \in \Omega$ let

$$\sum_{i,j=1}^{2} a_{i,j}\xi_i\xi_j \geqq \mu_0(\xi_1^2 + \xi_2^2)$$

with $\mu_0 > 0$.

Let $B(u, v)$ be a continuous bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$:

$$B(u, v) = \int_\Omega \left(\sum_{i,j=1}^{2} a_{i,j}u_{x_i}v_{x_j} + \sum_{i=1} b_i u_{x_i}v + cuv \right) dx.$$

We assume that

$$\inf_{\substack{\|u\|_{H^1(\Omega)}=1 \\ u \in H_0^1(\Omega)}} \sup_{\substack{\|v\|_{H^1(\Omega)}=1 \\ v \in H_0^1(\Omega)}} |B(u, v)| \geqq \mu_1 > 0$$

and for any $v \in H_0^1(\Omega)$, $v \neq 0$

$$\sup_{\substack{\|u\|_{H^1(\Omega)}=1 \\ u \in H_0^1(\Omega)}} |B(u, v)| > 0.$$

Assume now that $g^{[l]} \in \mathfrak{B}_\beta^{3/2-l}(\Gamma^{(l)})$, $l = 0, 1$, $f \in \mathfrak{B}_\beta^0(\Omega)$ and consider the boundary value problem

$$Lu = f \quad \text{on } \Omega, \tag{3.1a}$$

$$u = g^{[0]} \quad \text{on } \Gamma^{(0)}, \tag{3.1b}$$

$$\frac{\partial u}{\partial n_c} = g^{[1]} \quad \text{on } \Gamma^{(1)} \tag{3.1c}$$

where we denoted by n_c the conormal in the usual sense. The solution of our problem is understood in the usual sense. Then we have Theorem 3.1.

THEOREM 3.1. *There exists unique solution* $u_0 \in H^1(\Omega)$ *of the problem* (3.1). *See Lemma* 3.1 *of* [3].

Let us mention some theorems addressing regularity of the solution u_0.

THEOREM 3.2. *There exists* $0 \leqq \bar{\beta}_i < 1$, $i = 1, \cdots, M$, *depending on the problem* (*i.e., operator L,* ω_i, *etc.*), *such that if* $f \in \mathfrak{B}_\beta^0(\Omega)$, $g^{[l]} \in \mathfrak{B}_\beta^{3/2-l}(\Gamma^{(l)})$, $l = 0, 1$, $\bar{\beta} < \beta < 1$, *then* $u_0 \in \mathfrak{B}_\beta^2(\Omega)$.

Proof is given in [3].

THEOREM 3.3. *Let* Ω *be a* (*curvilinear*) *polygon* (*instead of straight polygon as in Theorem* 3.2) *and let the assumptions of Theorem* 3.2 *hold. Then* $u_0 \in \mathfrak{C}_\beta^2(\Omega)$.

Proof of the theorem is given in [4].

We have seen in [4], [5a], and [5b] that when the solution u of the problem (3.1a)-(3.1c) belongs to the class $\mathfrak{B}_\beta^2(\Omega)$ then the *h-p* version of the finite-element method converges exponentially.

Theorems 3.1 and 3.2 show that it is important to develop practical characterizations of spaces $\mathfrak{B}_\beta^{3/2-l}(\Gamma)$, $l = 0, 1$, which can be easily used in concrete cases to verify whether the input data, i.e., $g^{[l]}$ belong to the desired space. We will elaborate on this in the next section.

4. Traces and extensions of weighted Sobolev spaces. Characterization of the spaces $\mathfrak{B}_\beta^{3/2-l}(\Gamma)$. In this section we will elaborate on the characterization of the space $\mathfrak{B}_\beta^{3/2-l}(\Gamma)$, $l=0,1$, which leads to an easy verification in the concrete cases of applications.

LEMMA 4.1. *Let $\beta=(\beta_1,\beta_2)$, $0<\beta<\frac{1}{2}$ and $g\in H_\beta^{1,1}(I)$. Then*

(i) $g\in C^0(\bar I)$ *and* $\|g\|_{C^0(\bar I)}\leqq C\|g\|_{H_\beta^{1,1}(I)}$;

(ii) $|g(x)-g(-1)|\leqq C\hat\Phi_{1/2-\beta}(x)\|g\|_{H_\beta^{1,1}(I)}$,

$|g(x)-g(1)|\leqq C\hat\Phi_{1/2-\beta}(x)\|g\|_{H_\beta^{1,1}(I)}$,

where C is a constant independent of $g(x)$ (but depends on β).

Proof. Obviously,

$$
\begin{aligned}
|g(x)-g(t)| &\leqq \left|\int_t^x g'(\tau)\,d\tau\right| \\
&\leqq \left[\int_t^x g'^2(\tau)\hat\Phi_\beta^2(\tau)\,d\tau\right]^{1/2}\left[\int_t^x (\hat\Phi_\beta(\tau))^{-2}\,d\tau\right]^{1/2} \qquad (4.1)\\
&\leqq \|g\|_{H_\beta^{1,1}(I)}\left[\int_t^x (\hat\Phi_\beta(\tau))^{-2}\,d\tau\right]^{1/2},
\end{aligned}
$$

which shows that g is continuous on $\bar I$. Using the imbedding theorem on $(-\frac{1}{2},\frac{1}{2})=I'$, we have

$$
|g(0)|\leqq C\|g\|_{H^1(I')}\leqq C\|g\|_{H_\beta^{1,1}(I)} \qquad (4.2)
$$

and we get immediately

$$
\|g\|_{C^0(I)}\leqq C\|g\|_{H_\beta^{1,1}(I)}.
$$

Further, (4.1) immediately leads to (ii). □

LEMMA 4.2. *Let $\beta=(\beta_1,\beta_2)$, $\frac{1}{2}<\beta<1$ and $g\in H_\beta^{2,2}(I)$. Then*

(i) $g\in C^0(I)$ *and* $\|g\|_{C^0(I)}\leqq C\|g\|_{H_\beta^{2,2}(I)}$;

(ii) $|g(x)-g(-1)|\leqq C\hat\Phi_{3/2-\beta}(x)\|g\|_{H_\beta^{2,2}(I)}$,

$|g(x)-g(1)|\leqq C\hat\Phi_{3/2-\beta}(x)\|g\|_{H_\beta^{2,2}(I)}$,

where C is a constant independent of $g(x)$.

Proof. Using (4.1), we get

$$
\begin{aligned}
|g(x)-g(t)| &= \left|\int_t^x g'(\tau)\,d\tau\right| \\
&\leqq \left[\int_t^x g'^2(\tau)\hat\Phi_{1-\beta}^{-2}\,d\tau\right]^{1/2}\left[\int_t^x \hat\Phi_{1-\beta}^2(\tau)\,d\tau\right]^{1/2} \\
&\leqq \|g'\hat\Phi_{1-\beta}^{-1}\|_{L_2(I)}\left[\int_t^x \hat\Phi_{1-\beta}^2(\tau)\,d\tau\right]^{1/2}
\end{aligned}
$$

and

$$
\begin{aligned}
\|g'\hat\Phi_{1-\beta}^{-1}\|_{L_2(I)} &\leqq \|(g'-g'(0))\hat\Phi_{1-\beta}^{-1}\|_{L_2(I)}+|g'(0)|\,\|\hat\Phi_{1-\beta}^{-1}\|_{L_2(I)} \\
&\leqq C[|g'(0)|+\|g''\hat\Phi_\beta\|_{L_2(I)}] \\
&\leqq C\|g\|_{H_\beta^{2,2}(I)}.
\end{aligned}
$$

In the last inequality we used Lemma 2.5 and the fact that $\frac{1}{2}<\beta<1$. The lemma now follows immediately. □

LEMMA 4.3. *Let* $g \in \mathfrak{B}^1_{\beta,d}(I)$, $0<\beta<1$. *Then for* $k \geqq 1$

$$|g^{(k)}(x)| \leqq C(\hat{\Phi}_{k-1/2+\beta}(x))^{-1}(d_1)^k k!$$

where $d_1 = \gamma d$, $\gamma > 1$ *is independent of* g, k, d, *and* C *depends on* β, *but is independent of* g, k.

Proof. Let $I' = (-\frac{1}{2}, \frac{1}{2})$. *Then for any* $k \geqq 1$ we have

$$\|g^{(k)}\|_{H^1(I')} \leqq C(\hat{\Phi}(\tfrac{1}{2}))^{-k-\bar{\beta}} k! d^k$$

where $\bar{\beta} = \max(\beta_1, \beta_2)$. Hence by the imbedding theorem, we have

$$|g^{(k)}(0)| \leqq C d_1^k k!$$

where $d_1 \geqq \gamma d$, $\gamma > \hat{\Phi}^{-1}(\frac{1}{2}) > 1$. Further, for $k \geqq 1$, we have that

$$\begin{aligned}
|g^{(k)}(x)| &\leqq |g^{(k)}(0)| + \left| \int_0^x g^{(k+1)}(t)\, dt \right| \\
&\leqq |g^{(k)}(0)| + \left[\int_0^x (g^{(k+1)}(t))^2 \hat{\Phi}^2_{\beta+k}(t)\, dt \right]^{1/2} \left[\int_0^x \hat{\Phi}^{-2}_{\beta+k}(t)\, dt \right]^{1/2} \\
&\leqq C d_1^k k! [1 + \hat{\Phi}^{-1}_{\beta+k-1/2}(x)] \\
&\leqq C d_1^k k! (\hat{\Phi}_{k-1/2+\beta}(x))^{-1}.
\end{aligned}$$

□

COROLLARY 4.4. *Let* $g \in \mathfrak{B}^1_\beta(I)$, $0<\beta<1$. *Then* $g \in \mathfrak{C}^1_\beta(I)$.

COROLLARY 4.5. *Let* $g \in \mathfrak{B}^2_\beta(I)$, $0<\beta<1$. *Then for* $k \geqq 2$

$$|g^{(k)}(x)| \leqq C(\hat{\Phi}_{k-3/2+\beta}(x))^{-1} d_1^k k!$$

and $g \in \mathfrak{C}^2_\beta(I)$.

LEMMA 4.6. *Let* $\xi = m(x)$ *be a one-to-one map of* $\bar{I}$ *onto* $\bar{I}$, *let* $m(x)$ *be analytic on* $\bar{I}$, *and let* $|m'(x)| > 0$, $x \in \bar{I}$. *Assume that* $g \in \mathfrak{C}^j_\beta(I)$, $j = 1, 2$, *and define* $v(x) = g(m(x))$. *Then* $v \in \mathfrak{C}^j_\beta(I)$, $j = 1, 2$.

Proof. Because $m(x)$ is analytic on $\bar{I}$ it can be extended into the complex plane $\mathbb{C}$ on $I_\delta = \{z = x+iy \mid -1-\delta < x < 1+\delta, |y| < \delta\}$, $\delta > 0$, $m(z)$ is a one-to-one mapping of $\bar{I}_\delta$ onto $\bar{I}^*_\delta \supset I_{\delta'}$, $\delta' > 0$ and $|m'(z)| > \alpha_0 > 0$, $z \in \bar{I}_\delta$. Now let $j = 1$ and $x_0 \in I$. Then for $k \geqq 1$

$$|g^{(k)}(x_0)| \leqq C(\hat{\Phi}_{k-1/2+\beta}(x_0))^{-1} d_1^k k!$$

and the series

$$g'(x) = \sum_{k=0}^{\infty} g^{(k+1)}(x_0)(x-x_0)^k \frac{1}{k!}$$

is absolutely convergent for $|x - x_0| \leqq \alpha(\hat{\Phi}(x_0)/d_1)$, $\alpha < 1$. Hence also

$$g'(z) = \sum_{k=0}^{\infty} g^{(k+1)}(x_0)(z-x_0)^k \frac{1}{k}!$$

converges and $|g'(z)| \leqq C\hat{\Phi}^{-1}_{\beta+1/2}(x_0)$ for $|z - x_0| \leqq \alpha(\hat{\Phi}(x_0)/d_1)$ where C is independent of x_0. Hence $g(z)$ is a holomorphic function and $v(z) = g(m(z))$ is holomorphic, too. Using Cauchy's theorem we get immediately that for $k \geqq 1$

$$|v^{(k)}(x)| \leqq C d_2^k \hat{\Phi}^{-1}_{k-1/2+\beta}(x) k!.$$

Obviously, $v(x) \in H^{1,1}_\beta(I)$. In quite a similar way we prove the statement for $j = 2$. □

Remark 4.1. Lemma 4.6 shows that the space $\mathfrak{C}^j_\beta(I)$ is invariant with respect to an analytic mapping. Using the formula of the nth derivative of a composite function (see formula 0.430 of [8]) we can also show that $\mathfrak{B}^j_\beta(I)$ is an invariant space with respect to an analytic mapping $m(x)$ as in Lemma 4.6.

Let Γ be an analytic arc. Then we could define the spaces $\mathfrak{C}^j_\beta(\Gamma)$ and $\mathfrak{B}^j_\beta(\Gamma)$ with respect to the length instead as we did in § 2 by using a specific mapping. These two definitions are then equivalent by Lemma 4.6 and Remark 4.1.

LEMMA 4.7. *Let $M(x)$, $x\in\mathbb{R}^2$, $M(x)=(M_1(x), M_2(x))$ be a one-to-one mapping of $\bar\Omega$ onto $\bar{\tilde\Omega}$ and $|J^{-1}|\leqq\mu$ on $\bar\Omega$, where J is the Jacobian of the mapping. Assume that $M(x)$ can be analytically extended on $\bar\Omega_\delta=\{x\in\mathbb{R}^2 \mid \operatorname{dist}(x,\Omega)\leqq\delta\}$ so that it is a one-to-one mapping of $\bar\Omega_\delta$ onto $\bar\Omega^*$, $\Omega^*\supset\bar\Omega$. Let $u\in\mathfrak{C}^j_\beta(\Omega)$, $j=1,2$, $v(M(x))=u(x)$. Then $v\in\mathfrak{C}^j_\beta(\tilde\Omega)$.*

The proof is analogous to that of Lemma 4.6, however, we must apply the theory of two complex variables.

LEMMA 4.8. *Let $g\in\mathfrak{C}^j_\beta(I)$, $0<\beta<1$, $j=1,2$. Then*

$$g\in\mathfrak{B}^j_{\bar\beta}(I),\qquad 0<\bar\beta<1,$$

$$\bar\beta=\beta+\varepsilon,\quad \varepsilon>0 \quad \textit{arbitrary.}$$

Proof. Let us consider only the case $j=1$. The case $j=2$ is analogous. Because for $k\geqq 1$

$$|g^{(k)}(x)|\leqq Cd^k k!(\hat\Phi_{k+\beta-1/2}(x))^{-1}$$

we get

$$\int_{-1}^{1}(g^{(k)}(x))^2\hat\Phi^2_{k+\bar\beta-1}(x)\,dx\leqq Cd^{2k}(k!)^2\int_{-1}^{1}\hat\Phi^2_{\bar\beta-\beta-1/2}(x)\,dx$$

$$\leqq C(\varepsilon)d^{2k}(k!)^2. \qquad \square$$

We see that Lemma 2.3 has a completely analogous version for the relation between $\mathfrak{B}^2_\beta(I)$ and $\mathfrak{C}^2_\beta(I)$.

THEOREM 4.1. *Let $u\in H^{k+2,2}_\beta(\Omega)$, $k\geqq 0$, and Γ_i be a straight line edge of Ω and $u|_{\Gamma_i}=g_i$. Then we have the following*:

(i) *For $\frac12<\beta_i,\ \beta_{i+1}<1$ and $k\geqq 0$*

$$g_i\in H^{k+1,1}_{\hat\beta_i}(\Gamma_i),\qquad \hat\beta_i=(\hat\beta_{i,1},\hat\beta_{i,2}),$$

$$\hat\beta_{i,j}>0,\quad \hat\beta_{i,j}\in(\beta_{i+j-1}-\tfrac12,1),\quad j=1,2$$

and

$$\|g_i\|_{H^{k+1,1}_{\hat\beta_i}(\Gamma_i)}\leqq Cd^k\|u\|_{H^{k+2,2}_\beta(\Omega)}$$

with C independent of k and $d\geqq 1$.

(ii) *For $0<\beta_i,\ \beta_{i+1}<\frac12$, $k\geqq 1$*

$$g_i\in H^1(\Gamma_i),$$

$$g_i\in H^{k+1,2}_{\hat\beta_i}(\Gamma_i),\quad \hat\beta_{i,j}\in(\beta_{i+j-1}+\tfrac12,1),\quad j=1,2,$$

$$\|g_i\|_{H^1(\Gamma_i)}\leqq C\|u\|_{H^{2,2}_\beta(\Omega)},$$

$$\|g_i\|_{H^{k+1,2}_{\hat\beta_i}(\Gamma_i)}\leqq Cd^k\|u\|_{H^{k+2,2}_\beta(\Omega)}.$$

(iii) *If $u\in\mathfrak{B}^2_\beta(\Omega)$ and $\frac12<\beta_i,\ \beta_{i+1}<1$, then $g_i\in\mathfrak{B}^1_{\hat\beta_i}(\Gamma_i)$, $\hat\beta_{i,j}\in(\beta_{i+j-1}-\frac12,\frac12)$, $j=1,2$. If $0<\beta_i,\ \beta_{i+1}<\frac12$, then $g_i\in\mathfrak{B}^2_{\hat\beta_i}(\Gamma_i)$, $\hat\beta_{i,j}\in(\beta_{i+j-1}+\frac12,1)$.*

Proof. Without any loss of generality we can assume that $\Gamma_i = \Gamma_1$ and

$$\Gamma_1 = \{x_1, x_2 | x_1 \in I, x_2 = 0\}, \quad A_1 = (-1, 0), \quad A_2 = (1, 0), \quad \tilde{\beta} = (\beta_1, \beta_2).$$

Let $k \geqq 0$ and $v_k = (\partial^k u / \partial x_1^k)\Phi_{k+\beta}$. Then for $k \geqq 2$,

$$\begin{aligned}\|v_k\|_{H^2(\Omega)} &\leqq C\left[\left\|\left(D^2 \frac{\partial^k u}{\partial x_1^k}\right)\Phi_{k+\beta}\right\|_{L_2(\Omega)} + k\left\|\left(D^1 \frac{\partial^k u}{\partial x_1^k}\right)\Phi_{k+\beta-1}\right\|_{L_2(\Omega)}\right. \\ &\quad \left. + k^2\left\|\left(\frac{\partial^k u}{\partial x_1^k}\right)\Phi_{k+\beta-2}\right\|_{L_2(\Omega)}\right] \\ &\leqq Ck^2\|u\|_{H_\beta^{k+2,2}(\Omega)}.\end{aligned}$$

Using Lemma 2.2, we get for $k = 1$

$$|v_1|_{H^2(\Omega)} \leqq C\|u\|_{H_\beta^{3,2}(\Omega)}.$$

Because of Lemma 2.1 $u \in C^0(\bar{\Omega})$, and hence $v_0(A_i) = 0$, $i = 1, 2$. Hence, using Lemma 2.2, we get

$$|v_0|_{H^2(\Omega)} \leqq C\|u\|_{H_\beta^{2,2}(\Omega)},$$

and hence for all $k \geqq 0$

$$\|v_k\|_{H^2(\Omega)} \leqq C(k+1)^2\|u\|_{H_\beta^{k+2,2}(\Omega)} \tag{4.3}$$

where C is independent of k. Therefore by the imbedding theorem

$$v_k \in C^0(\bar{\Omega}), \qquad k \geqq 1.$$

Let us now show that $v_k(A_i) = 0$, $i = 1, 2$, $k \geqq 1$. Assume on the contrary that $v_k^2(A_1) > 0$. Then because $v_k \in C^0(\bar{\Omega})$ we have

$$v_k^2(x) > \varepsilon > 0 \quad \text{for } |x - A_1| < \delta, \quad \delta > 0.$$

Hence for $k \geqq 2$

$$\begin{aligned}\infty > \int_\Omega \Phi_{k+\beta-2}^2 \left(\frac{\partial^k u}{\partial x_1^k}\right)^2 dx &= \int_\Omega \Phi_{-2}^2 v_k^2 \, dx \\ &\geqq \varepsilon^2 \int_{\Omega_\delta'} \Phi_{-2}^2 \, dx = \infty\end{aligned}$$

where

$$\Omega_\delta' = \Omega \cap \{x \,|\, |x - A_1| < \delta\}$$

and we have the desired contradiction. For $k = 1$ we use Lemma 2.2 and get

$$\infty > \int_\Omega \left(\frac{\partial u}{\partial x_1}\right)^2 \Phi_{\beta-1}^2 \, dx > \varepsilon^2 \int_{\Omega_\delta'} \Phi_{-2}^2 \, dx = \infty.$$

If $u \in \mathfrak{B}_\beta^2(\Omega)$ then we get from (4.3) for $k \geqq 0$

$$\|v_k\|_{H^2(\Omega)} \leqq Cd_1^k k!.$$

We have $g_1^{(k)}(x_1) = \partial^k u / \partial x_1^k|_{\Gamma_1}$, $k \geqq 0$. Then $g_1^{(k)}(x_1) = \Phi_{k+\beta}^{-1}(x)v_k(x)|_{\Gamma_1} = \Phi_{k+\beta}^{-1}(x_1)v_k(x_1)$ where we wrote $\Phi_{k+\beta}^{-1}(x_1)$ and $v_k(x_1)$ instead of $\Phi_{k+\beta}^{-1}(x_1, 0)$ and $v_k(x_1, 0)$. Assume first that $\frac{1}{2} < \beta_1, \beta_2 < 1$. Let $d_0 = \{\min_{j=3,\cdots,M} \operatorname{dist}(A_j, \Gamma_1)\}^{M-2}$. Then we have for $x \in \Gamma_1$, $\hat{\Phi}(x_1) \leqq \Phi(x_1)d_0^{-1}$, and hence for $j = 1, 2, \cdots, k+1$,

$$\begin{aligned}\int_{\Gamma_1} \hat{\Phi}_{j-1+\hat{\beta}_1}^2 |g_1^{(j)}(x_1)|^2 \, dx_1 &\leqq Cj^2\left\{\int_{-1}^1 \hat{\Phi}_{j-1+\hat{\beta}_1}^2 [|v_{j-1}'|^2 \Phi_{\beta+j-1}^{-2} + \Phi_{j+\beta}^{-2}|v_{j-1}|^2] \, dx_1\right\} \\ &\leqq Cd_0^{-2j}j^2\left\{\int_{-1}^1 [|v_{j-1}'|^2 \hat{\Phi}_{\hat{\beta}_1-\tilde{\beta}}^2 + |v_{j-1}|^2 \hat{\Phi}_{-1+\hat{\beta}_1-\tilde{\beta}}^2] \, dx_1\right\}.\end{aligned}$$

Using Lemma 2.5, the fact that $j=1,\cdots,k+1$, $v_{j-1}(A_i)=0$, $i=1,2$ and that $\tilde{\beta}-\hat{\beta}_1+1>\frac{1}{2}$, we get for some $d_1<1$

$$\int_{\Gamma_1}\hat{\Phi}^2_{j-1+\hat{\beta}}|g_1^{(j)}(x_1)|^2\,dx_1\leqq Cd_1^{-2j}\int_{-1}^{1}|v'_{j-1}|^2\hat{\Phi}^2_{\hat{\beta}_1-\tilde{\beta}}\,dx_1.$$

By (4.3) and the imbedding theorem we have for $1<p<\infty$ and $j=1,\cdots,k+1$

$$\|v'_{j-1}\|_{L_p(I)}\leqq C(p)\|v_{j-1}\|_{H^2(\Omega)}\leqq Cj^2\|u\|_{H^{j+1,2}_{\beta}(\Omega)}.$$

Hence for $j=1,\cdots,k+1$, because $\hat{\beta}_1-\tilde{\beta}>-\frac{1}{2}$, we get

$$\begin{aligned}\int_{-1}^{1}\hat{\Phi}^2_{j-1+\hat{\beta}_1}|g^{(j)}(x_1)|^2\,dx&\leqq Cd_1^{-2k}\left(\int_{-1}^{1}(\Phi^2_{\hat{\beta}_1-\tilde{\beta}})^p\,dx\right)^{1/p}\|v'_{j-1}\|^2_{L_{2q}(I)}\\&\leqq Cd_1^{-2k}\|v_{j-1}\|^2_{H^2(\Omega)}\\&\leqq Cd_2^{-2k}\|u\|^2_{H^{k+2,2}_{\beta}(\Omega)}.\end{aligned}$$

Because by Lemma 2.1

$$\|u\|_{C^0(\bar{\Omega})}\leqq C\|u\|_{H^{2,2}_{\tilde{\beta}}(\Omega)},$$

we get

$$\|g\|_{L_2(\Gamma_1)}\leqq C\|u\|_{H^{2,2}_{\tilde{\beta}}(\Omega)}.$$

Hence we have proven (i) and (iii) for $\frac{1}{2}<\beta_i,\beta_{i+1}<1$ and $k\geqq 0$. Assume now that $0<\beta_1,\beta_2<\frac{1}{2}$. We will proceed analogously as before. For $j\geqq 2$, we have

$$\begin{aligned}\int_{\Gamma_1}\hat{\Phi}^2_{j-2+\hat{\beta}_1}|g^{(j)}(x_1)|^2\,dx_1&\leqq Cj^2\int_{-1}^{1}\hat{\Phi}^2_{j-2+\hat{\beta}_1}[|v'_{j-1}|^2\Phi^{-2}_{\beta+j-1}+\Phi^{-2}_{j+\beta}|v_{j-1}|^2]\,dx_1\\&\leqq Cd_1^{-2j}\int_{-1}^{1}\hat{\Phi}^2_{-1+\hat{\beta}_1-\tilde{\beta}}(v'_{j-1}(x_1))^2\,dx_1\end{aligned}$$

where we once used Lemma 2.5 and the fact that $-1+\hat{\beta}_1-\tilde{\beta}>-\frac{1}{2}$. Hence, using (4.3) and realizing that $-1+\hat{\beta}_1-\tilde{\beta}>-\frac{1}{2}$, we get analogously as before for $j=2,\cdots,k+1$

$$\int_{\Gamma_1}\hat{\Phi}^2_{j-2+\hat{\beta}_1}|g^{(j)}(x_1)|^2\,dx_1\leqq Cd_2^{2k}\|u\|^2_{H_2^{k+2,2}(\Omega)}.$$

Let us prove now that

$$\|g\|_{H^1(\Gamma_i)}\leqq C\|u\|_{H^{k+2,2}_{\beta}(\Omega)},\qquad k\geqq 1.$$

We have $v_0(A_1)=v_0(A_2)=0$, and hence

$$\int_{-1}^{1}g'^2\,dx\leqq Cd_0^{-2}\int_{-1}^{1}[\hat{\Phi}^{-2}_{\tilde{\beta}}|v'_0|^2+|v_0|^2\hat{\Phi}^{-2}_{\tilde{\beta}+1}]\,dx\leqq\tilde{C}d_0^{-2}\int_{-1}^{1}\hat{\Phi}^{-2}_{\tilde{\beta}}|v'_0|^2\,dx$$

where we have again used Lemma 2.5. Because $0<\tilde{\beta}<\frac{1}{2}$ and

$$\|v'_0\|_{L_p(I)}\leqq C(p)\|v_0\|_{H^2(\Omega)}\leqq C\|u\|_{H^{2,2}_{\tilde{\beta}}(\Omega)},$$

we proceed as before and (ii) and (iii) follow easily. □

Remark 4.2. In the proof of Theorem 4.1 it has been assumed that $\hat{\beta}_{i,j}\in(\beta_{i+j-1}-\frac{1}{2},1)$, respectively, $\hat{\beta}_{i,j}\in(\beta_{i+j-1}+\frac{1}{2},1)$, i.e., of the open interval. The proof does not hold for the closed interval. It has been assumed in Lemma 4.9 that the edge Γ_i of the domain was straight. Let us now assume that $\Gamma_i=m(\bar{I})$ where $m=(\varphi,\psi)$ are analytic functions on I as given in § 2. Then we have Lemma 4.9.

LEMMA 4.9. *Let the edge Γ_i of the domain be analytic. Then part* (iii) *of Theorem* 4.1 *holds.*

Proof. By Lemma 2.3, $u \in \mathfrak{C}^2_\beta(\Omega)$. Let $M(\xi) = (\varphi(\xi), \psi(\xi))$, $\xi \in I$, be the mapping of I onto Γ_1. Then we define

$$M_1(\xi, \eta) = \varphi(\xi) - \eta\psi'(\xi), \qquad M_2(\xi, \eta) = \psi(\xi) + \xi\varphi'(\xi).$$

Then the mapping $M(\xi, \eta) = (M_1(\xi, \eta), M_2(\xi, \eta))$ is analytic on $I_\delta = \{\xi, \eta \,|\, -1-\delta < \xi < 1+\delta, |\eta| < \delta\}$, $\delta > 0$, $|J| < \alpha$, $|J^{-1}| < \alpha$ on I_δ (where J is the Jacobian of the mapping) and maps I_δ onto the (open) neighborhood S^* of Γ_i. Denoting $\Omega^* = \Omega \cap S^*$, $T = M^{-1}(\Omega^*)$, we see that $v(x) = u(M^{-1}(x))$ is defined on T, and $v \in \mathfrak{C}^2_\beta(T)$ by using Lemma 4.7. Hence $v \in \mathfrak{B}^2_{\beta+\varepsilon}(T)$, $\varepsilon > 0$ arbitrary, by Lemma 2.3. Hence for $\frac{1}{2} < \beta_i, \beta_{i+1} < 1$ we get by Theorem 4.1(iii)

$$g_i(\xi) = v(\xi, 0) \in \mathfrak{B}^1_{\hat\beta_i}(I_i), \quad \hat\beta_{i,j} \in (\beta_{i+j-1} + \varepsilon - \tfrac{1}{2}, \tfrac{1}{2}), \quad j = 1, 2.$$

Because $\varepsilon > 0$ is arbitrary, $\hat\beta_{i,j} \in (\beta_{i+j-1} - \frac{1}{2}, \frac{1}{2})$. Analogously for $0 < \beta_i, \beta_{i+1} < \frac{1}{2}$, $g_i(\xi) \in \mathfrak{B}^2_{\hat\beta_i}(I)$, $\hat\beta_{i,j} \in (\beta_{i+j-1} + \frac{1}{2}, 1)$. □

LEMMA 4.10. *Let $g_1 \in \mathfrak{B}^1_\beta(I)$, $0 < \hat\beta_1 < \frac{1}{2}$, $0 < \hat\beta_2 < 1$, $g_2 \in \mathfrak{B}^2_{\hat\beta}(I)$, $\frac{1}{2} < \hat\beta_1 < 1$, $0 < \hat\beta_2 < 1$. Let $S = \{r, \theta \,|\, 0 < \theta < 2\pi, 0 < r < 1\}$ where (r, θ) are polar coordinates with respect to $(-1, 0)$ and $\Phi(r) = r$. Define*

$$U_i(r, \theta) = g_i(-1 + r), \qquad V_i(r, \theta) = \theta[g_i(-1 + r) - g_i(-1)]$$

(by Theorems 3.1, 3.2, *$g_i \in C^0(\bar I)$, $i = 1, 2$, and hence $g_i(-1)$ is well defined). Then*

$$U_1, V_1 \in \mathfrak{B}^2_\beta(S), \qquad \beta = \hat\beta_1 + \tfrac{1}{2},$$

$$U_2, V_2 \in \mathfrak{B}^2_\beta(S), \qquad \beta = \tilde\beta_1 - \tfrac{1}{2}.$$

Proof. Assume first that $0 < \hat\beta_1 < \frac{1}{2}$ and $g_1 \in \mathfrak{B}^1_\beta(I)$. Set $\beta = \hat\beta_1 + \frac{1}{2}$ and $U_1 = g_1(-1 + r)$. Then for $k \geqq 2$

$$\int_S \left(\frac{\partial^k U_1}{\partial r^k}\right)^2 (r^{k-2+\beta})^2 \, r \, dr \, d\theta \leqq Cd^{2k} \|g_1^{(k)} \hat\Phi_{k-1+\hat\beta}\|^2_{L_2(I)}$$

$$\leqq Cd_1^{2k}(k!)^2.$$

Hence by Theorem 1.1 of [4] we have for $k \geqq 2$, $|\alpha| = k$

$$\| |D^\alpha U_1| \phi_{\beta+k-2} \|_{L_2(S)} \leqq Cd_2^k k! \, .$$

Furthermore,

$$\| U_1 \|_{H^1(S)} \leqq C \| g_1^{(1)} \Phi_{1/2} \|_{L_2(I)}$$

$$\leqq C \| g_1^{(1)} \hat\Phi_{\hat\beta_1} \|_{L_2(I)}.$$

Hence, $U_1 \in \mathfrak{B}^2_\beta(S)$. Now let $\frac{1}{2} < \tilde\beta_1 < 1$. Set $\beta = \tilde\beta_1 - \frac{1}{2}$. As before, we have for $k \geqq 2$

$$\int_S \left(\frac{\partial^k U_2}{\partial r^k}\right)^2 (r^{k-2+\beta})^2 r \, dr \, d\theta \leqq Cd_1^{2k}(k!)^2$$

and we get $\| U_2 \|_{H^1(S)} < \infty$. Hence, $U_2 \in \mathfrak{B}^2_\beta(S)$. Let us now consider the function $V_1(r, \theta)$. Then as before

$$\int_S \left(\frac{\partial^k V_1}{\partial r^k}\right)^2 (r^{k-2+\beta})^2 r \, dr \, d\theta \leqq Cd_1^{2k}(k!)^2.$$

Furthermore, using Lemma 2.5 and $k \geqq 2$, we get

$$\begin{aligned}\int_S \left(\frac{\partial^k V_1}{\partial r^{k-1}\,\partial\theta}\right)^2 r^{-2}(r^{k-2+\beta})^2 r\,dr\,d\theta &= \int_S \left(\frac{\partial^{k-1} g_1}{\partial r^{k-1}}\right)^2 r^{-2}(r^{k-2+\beta})^2 r\,dr\,d\theta \\ &\leqq Cd_1^{2k}\|g_1^{(k-1)}\hat{\Phi}_{k-2+\hat{\beta}_1}\|^2_{L_2(I)} \\ &\leqq Cd^{2k}[\|(g_1^{(k-1)} - g_1^{(k-1)}(0))\hat{\Phi}_{k-2+\hat{\beta}_1}\|^2_{L_2(I)} \\ &\qquad + (g_1^{(k-1)}(0))^2\|\hat{\Phi}_{k-2+\hat{\beta}_1}\|^2_{L_2(I)}] \\ &\leqq Cd_2^{2k}[(g_1^{(k-1)}(0))^2 + \|g_1^{(k)}\hat{\Phi}_{k-1+\hat{\beta}_1}\|^2_{L_2(I)}] \\ &\leqq Cd_3^{2k}(k!)^2.\end{aligned}$$

In the last inequality we used the fact that

$$|g^{(k-1)}(0)| \leqq Cd_4^k(k!)$$

and realizing that $\partial^k V_1/(\partial r^{k-j}\,d\theta^j) = 0$ for $k \geqq j \geqq 2$ we have for $|\alpha| = k \geqq 2$

$$\|\,|D^\alpha V_1|\Phi_{\beta+k-2}\|_{L_2(S)} \leqq Cd^k k!\,.$$

Furthermore for $0 < \hat{\beta}_1 < \frac{1}{2}$ and $I^* = (-1, 0)$ we have

$$\begin{aligned}\|V_1\|^2_{H^1(S)} &\leqq C[\|g_1^{(1)}\hat{\Phi}_{1/2}\|^2_{L_2(I^*)} + \|(g_1(x) - g_1(-1))\hat{\Phi}_{-1/2}\|^2_{L_2(I^*)}] \\ &\leqq C[\|g_1^{(1)}\hat{\Phi}_{\hat{\beta}_1}\|^2_{L_2(I^*)} + \|(g_1(x) - g_1(-1))\hat{\Phi}_{-1+\hat{\beta}_1}\|^2_{L_2(I^*)}] \\ &\leqq C[\|g_1^{(1)}\hat{\Phi}_{\hat{\beta}_1}\|^2_{L_2(I^*)} + \|g_1^{(1)}\hat{\Phi}_{\hat{\beta}_1}\|^2_{L_2(I)}] \\ &\leqq C\|g_1\|^2_{H^{1,1}_{\hat{\beta}}(I)}.\end{aligned}$$

In the last inequality we have used once more Lemma 2.5 and the fact that $\hat{\beta}_1 < \frac{1}{2}$. Quite analogously we prove that $V_2 \in \mathfrak{B}^2_\beta(S)$. □

LEMMA 4.11. *Let* $g \in \mathfrak{B}^1_{\hat{\beta}}(I)$, $0 < \hat{\beta} < \frac{1}{2}$, $g(\pm 1) = 0$. *Then for* $0 < \gamma < \frac{1}{2}$, $v = g\hat{\Phi}_{-\gamma} \in \mathfrak{B}^1_{\hat{\beta}+\gamma}(I)$.

Proof. For $k \geqq 1$,

$$\begin{aligned}&\int_{-1}^1 ((v^{(k)})^2\hat{\Phi}^2_{k-1+\hat{\beta}+\gamma})\,dx \\ &\quad\leqq \int_{-1}^1 \left(\sum_{l=0}^k \binom{k}{l} g^{(l)}(\hat{\Phi}_{-\gamma})^{(k-l)}\right)^2 \hat{\Phi}^2_{k-1+\hat{\beta}+\gamma}\,dx \\ &\quad\leqq Cd^{2k}\sum_{l=0}^k \int_{-1}^1 (g^{(l)})^2((k-l)!)^2\hat{\Phi}^2_{-\gamma-(k-l)}\hat{\Phi}^2_{k-1+\hat{\beta}+\gamma}\,dx \\ &\quad\leqq Cd^{2k}\left[\sum_{l=1}^k \int_{-1}^1 (g^{(l)})^2\hat{\Phi}^2_{\hat{\beta}+l-1}((k-l)!)^2\,dx + \int_{-1}^1 (g)^2\hat{\Phi}^2_{\hat{\beta}-1}(k!)^2\,dx\right] \\ &\quad\leqq Cd^{2k}\left[\sum_{l=1}^k \int_{-1}^1 (g^{(l)})^2\hat{\Phi}_{\hat{\beta}+l-1}((k-l)!)^2\,dx + \int_{-1}^1 (g')^2\hat{\Phi}^2_{\hat{\beta}}(k!)^2\,dx\right] \leqq Cd_1^{2k}(k!)^2\end{aligned}$$

where we have used Lemma 2.5 in the above inequality. Furthermore,

$$\int_{-1}^1 v^2\,dx = \int_{-1}^1 g^2\Phi^2_{-\gamma}\,dx \leqq C\|g\|^2_{H^{1,1}_{\hat{\beta}}(I)}$$

by Lemma 4.1. □

LEMMA 4.12. *Let* $g \in \mathfrak{B}^2_{\hat\beta}(I)$, $g(\pm 1)=0, \frac{1}{2}<\hat\beta<1$, $0<\gamma<\frac{1}{2}$, $v=g\hat\Phi_{-\gamma}$. *Then for* $\hat\beta+\gamma>1$, $v\in\mathfrak{B}^1_{\hat\beta+\gamma-1}(I)$ *and for* $\hat\beta+\gamma<1$, $v\in\mathfrak{B}^2_{\hat\beta+\gamma}(I)$.

Proof. (a) Assume first that $\hat\beta+\gamma>1$. Then for $k\geqq 2$

$$\int_{-1}^{1}(v^{(k)})^2\hat\Phi^2_{k+\hat\beta+\gamma-2}\,dx \leqq Cd^{2k}\Bigg[\sum_{l=2}^{k}\int_{-1}^{1}(g^{(l)})^2\hat\Phi^2_{-\gamma-(k-l)+k+\hat\beta+\gamma-2}((k-l)!)^2\,dx + (k!)^2\int_{-1}^{1}g^2\hat\Phi^2_{\hat\beta-2}\,dx+((k-1)!)^2\int_{-1}^{1}g'^2\hat\Phi^2_{\hat\beta-1}\,dx\Bigg]$$

$$\leqq Cd^{2k}\Bigg[\sum_{l=2}^{k}\int_{-1}^{1}(g^{(l)})^2\hat\Phi^2_{\hat\beta+l-2}((k-l)!)^2\,dx + (k!)^2\int_{-1}^{1}g'^2\hat\Phi^2_{\hat\beta-1}\,dx\Bigg].$$

In the last inequality, Lemma 2.5 has been used. Because by the imbedding theorem $|g'(0)|\leqq C\|g\|_{H^{2,2}_{\hat\beta}(I)}$ and using Lemma 2.5 once more yields $\hat\beta-1>-\frac{1}{2}$, we get

$$\int_{-1}^{1}g'^2\hat\Phi^2_{\hat\beta-1}\,dx\leqq C\Bigg[\int_{-1}^{1}g''^2\hat\Phi^2_{\hat\beta}\,dx+|g'(0)|^2\Bigg]\leqq C\|g\|^2_{H^{2,2}_{\hat\beta}(I)}<\infty.$$

Hence,

$$\int_{-1}^{1}(v^{(k)})^2\hat\Phi_{k+\hat\beta+\gamma-2}\,dx\leqq Cd_1^{2k}(k!)^2.$$

Furthermore, as before

$$\int_{-1}^{1}v'^2\hat\Phi^2_{\hat\beta+\gamma-1}\,dx\leqq C\int_{-1}^{1}g'^2\hat\Phi^2_{\hat\beta-1}\,dx\leqq C\|g\|^2_{H^{2,2}_{\hat\beta}(I)}<\infty.$$

Because $g\in C^0(\bar I)$, $v\in L_2(I)$.

(b) Now assume that $\hat\beta+\gamma<1$. Then for $k\geqq 2$ we get exactly as before that

$$\int_{-1}^{1}(v^{(k)})^2\hat\Phi^2_{k+\hat\beta+\gamma-2}\,dx\leqq Cd_1^{2k}(k!)^2.$$

Furthermore,

$$\int_{-1}^{1}v'^2\,dx\leqq C\Bigg[\int_{-1}^{1}g^2\hat\Phi^2_{-\gamma-1}\,dx+\int_{-1}^{1}g'^2\hat\Phi^2_{-\gamma}\,dx\Bigg]$$

$$\leqq C\Bigg[\int_{-1}^{1}g''^2\hat\Phi^2_{-\gamma+1}\,dx+|g'(0)|^2\Bigg].$$

Because $-\gamma+1>\hat\beta$ by our assumption we see that

$$\int_{-1}^{1}v'^2\,dx\leqq C\|g\|^2_{H^{2,2}_{\hat\beta}(I)}.$$

Using Lemma 4.2, we also get $\|v\|_{L_2(I)}\leqq C\|g\|_{H^{2,2}_{\hat\beta}(I)}$. □

LEMMA 4.13. *Let* Ω *be a curvilinear polygon with the vertices* A_i, $i=1,\cdots,M$. *Let* $u\in\mathfrak{B}^2_\beta(\Omega)$ *and* w *be such that*

$$|D^\alpha w|\leqq C\Phi_{-|\alpha|+\gamma}|\alpha|!d^{|\alpha|},$$

$$\gamma=(\gamma_1,\cdots,\gamma_M),\quad |\alpha|\geqq 0,\quad \beta_i-\gamma_i>0,\quad \gamma_i\geqq 0.$$

Then $v=wu\in\mathfrak{B}^2_{\bar\beta}(\Omega)$ *where* $\bar\beta_i=\beta_i-\gamma_i$.

Proof. For $k \geqq 2$, $|\alpha| = k$, we have

$$\int_\Omega |D^{|\alpha|}v|^2 \Phi^2_{|\alpha|-2+\bar\beta}\, dx \leqq Cd^{2k}\left(\sum_{l=0}^{k}\int_\Omega |D^{k-l}u||D^l w|\right)^2 \Phi^2_{k-2+\bar\beta}\, dx$$

$$\leqq Cd_1^{2k}\sum_{l=0}^{k}((l+1)!)^2\int_\Omega |D^{k-l}u|^2\Phi^2_{k-2-l+\beta}\, dx$$

$$\leqq Cd_1^{2k}\sum_{l=0}^{k}((l+1)!)^2((k-1+l)!)^2 \leqq Cd_2^{2k-2}((k-2)!)^2.$$

Furthermore,

$$\int_\Omega |D^1 v|^2\, dx \leqq C\left[\int_\Omega |D^1 u|^2|w|^2\, dx + \int_\Omega |u|^2|D^1 w|^2\, dx\right] < \infty$$

because by Lemma 2.1 $u \in C^0(\bar\Omega)$. □

It is very easy to prove the following lemma.

LEMMA 4.14. *Let* $g \in \mathfrak{B}^0_{\hat\beta}(I)$, $0 < \hat\beta < \frac{1}{2}$. *Then* $v = g\hat\Phi \in \mathfrak{B}^1_{\hat\beta}(I)$ *and* $v(\pm 1) = 0$. *Let* $g \in \mathfrak{B}^1_{\hat\beta}(I)$, $\frac{1}{2} < \hat\beta < 1$; *then* $v = g\hat\Phi \in \mathfrak{B}^2_{\hat\beta}(I)$ *and* $v(\pm 1) = 0$.

Proof. The statement that $v \in \mathfrak{B}^1_{\hat\beta}(I)$ can be directly verified. By Lemma 4.1 v is continuous on $\bar I$. If $v(-1) \neq 0$, then $v^2(x) > \varepsilon > 0$ for all $|x+1| < \delta$. Hence, $g^2 = (v\hat\Phi^{-1})^2 \geqq \varepsilon\hat\Phi^2_{-1}$, which contradicts the assumption that $g \in \mathfrak{B}^0_{\hat\beta}(I)$, $0 < \hat\beta < \frac{1}{2}$. The proof of the second part of the lemma is analogous. □

LEMMA 4.15. *Let* $u \in \mathfrak{B}^2_\beta(\Omega)$, $0 < \beta < 1$ *and* $u = 0$ *at* A_i. *Then* $u\Phi^{-1} \in \mathfrak{B}^1_\beta(\Omega)$.

The proof follows easily using Lemma 2.2.

THEOREM 4.2. *Let* Ω *be a straight polygon with the edges* Γ_i, $i = 1, \cdots, M$, *and let* $g \in \mathfrak{B}^1_{\hat\beta}(\Gamma_1)$, $0 < \hat\beta_i < \frac{1}{2}$, $\beta_i = \hat\beta_i + \frac{1}{2}$, $i = 1, 2$ (*respectively,* $g \in \mathfrak{B}^2_{\hat\beta}(\Gamma_1)$, $\frac{1}{2} < \hat\beta_i < 1$, $\beta_i = \hat\beta_i - \frac{1}{2}$, $i = 1, 2$) *and* $g(A_i) = 0$, $i = 1, 2$. *Then there is* u *such that*

(i) $u \in \mathfrak{B}^2_\beta(\Omega)$, *with* $0 < \beta_j < 1$, $j = 3, \cdots, M$, *arbitrary*;

(ii) $u|_{\Gamma_1} = g$ *and* $u|_{\Gamma_j} = 0$ *for* $j = 2, \cdots, M$.

Proof. Let $\psi = \prod_{i=3}^{M}|x - A_i|^2$, $x \in \Omega$. Denote $\tilde g = g/\psi$. Then obviously $\tilde g \in \mathfrak{B}^1_{\hat\beta}(\Gamma_1)$ (respectively, $\tilde g \in \mathfrak{B}^2_{\hat\beta}(\Gamma_1)$). Now select $0 < \gamma_i < \frac{1}{2}$ such that $0 < \hat\beta + \gamma_i < \frac{1}{2}$ (respectively, $0 < \hat\beta + \gamma_i - 1 < \frac{1}{2}$). Denote $\hat g = \tilde g \prod_{i=1}^{2}|x - A_i|^{-\gamma_i} = \tilde g\Phi_{-\gamma}$ where $\gamma = (\gamma_1, \gamma_2, 0, \cdots, 0)$. By Lemmas 4.1 and 4.2 $\hat g(A_i) = 0$, $i = 1, 2$. Using Lemma 4.11 (and 4.12) we see that $\hat g \in \mathfrak{B}^1_{\hat\beta+\gamma}(I)$ (respectively, $\hat g \in \mathfrak{B}^1_{\hat\beta+\gamma-1}(I)$).

Let $U \in H^1(\Omega)$, $\Delta U = 0$ and $U = \hat g$ on Γ_1 and $U = 0$ on Γ_j, $j = 2, \cdots, M$. Function U exists and is uniquely determined. To see this consider $\varphi(x)$, $x \in \Gamma_1$, $\varphi \in C^\infty(\Gamma_1)$, $\varphi(x) = 1$ for $|x - A_i| \leqq \varepsilon/2$, $i = 1, 2$ and $\varphi(x) = 0$ for $|x - A_i| > \varepsilon$, $i = 1, 2$ with ε sufficiently small. We define

$$U = U_1 + U_2$$

where $\Delta U_i = 0$, $U_i \in H^1(\Omega)$, $i = 1, 2$, $U_1|_{\Gamma_1} = \hat g(1 - \varphi)$, $U_2|_{\Gamma_1} = \hat g\varphi$ and $U_i = 0$ on Γ_j, $j = 2, \cdots, M$. Because $h_1 = \hat g(1 - \varphi) \in C^\infty(\Gamma_1)$ and $h_1(x) = 0$ for $|x - A_i| \leqq \varepsilon/2$, U_1 obviously exists.

By Lemma 4.10 there exists $W \in H^1(\Omega)$ such that $W|_{\Gamma_1} = h_2 = \hat g\varphi$, and $W|_{\Gamma_j} = 0$, $j = 2, \cdots, M$. Hence, U_2 exists too. Function U has the following properties:

(i) $\Delta U = 0$.

(ii) $U|_{\Gamma_1} = \hat g$, $U|_{\Gamma_j} = 0$, $j = 2, \cdots, M$.

(iii) $\hat g$ is analytic on Γ_1 (not on $\bar\Gamma_1$).

(iv) In $\Omega_{i,\delta} = \Omega \cap \{x \,|\, |x - A_i| < \delta\}$, $i = 1, 2$, with δ sufficiently small, there is W_i such that $W_i \in \mathfrak{B}^2_{\bar\beta}(\Omega_{i,\delta})$, where $\bar\beta = \hat\beta + \gamma + \frac{1}{2}$ (respectively, $\bar\beta = \hat\beta + \gamma - 1 + \frac{1}{2}$) and $W_i|_{\Gamma_i \cap \bar\Omega_{i,\delta}} = \hat g$. (This follows from Lemma 4.10.)

By the selection of γ_i we have $\bar{\beta}_i > \frac{1}{2}$, $i = 1, 2$. Now using the same arguments as in the proof of Theorem 2.1 in [4], we conclude that $U \in \mathfrak{B}^2_{\bar{\beta}}(\Omega)$, where $\bar{\beta}_i = \hat{\beta}_i + \gamma_i + \frac{1}{2}$ (respectively, $\bar{\beta}_i = \hat{\beta}_1 + \gamma_i - \frac{1}{2}$), $i = 1, 2$, and $1 > \bar{\beta}_j > \frac{1}{2}$.

By Lemma 4.13 we see that $u = \psi \Phi_\gamma U \in \mathfrak{B}^2_\beta(\Omega)$ where $\beta_i = \hat{\beta}_i + \frac{1}{2}$ (respectively, $\beta_i = \hat{\beta}_i - \frac{1}{2}$), $i = 1, 2$ and $0 < \beta_j < 1$ arbitrary for $j = 3, \cdots, M$. In addition, $u|_{\Gamma_1} = g$ and $u|_{\Gamma_j} = 0$, $j = 2, \cdots, M$.

Let us outline the main idea of the assertion that $U \in \mathfrak{B}^2_{\bar{\beta}}(\Omega)$. Let $S_{i,\delta_i} = \{r_i, \theta_i | 0 < r_i < \delta_i, 0 < \theta_i < \omega_i\} \cap \Omega$ where (r_i, θ_i) are the polar coordinates with the origin in A_i. We select $\delta_i < 1$ such that $S_{i,2\delta_i} \cap S_{j,2\delta_j} = \varnothing$ for $i \neqq j$. Using Theorems 5.7.1, 5.7.1′, and 6.6.1 of [8], we conclude similarly, as in the proof of Theorem 2.1 of [3], that $U \in \mathfrak{B}^2_{\bar{\beta}}(\Omega - \bigcup_{i=1}^M S_{i,\delta_i/4})$ due to the analyticity of $\hat{g}$ on $\Gamma - \bigcup_{i=1}^M S_{i,\delta_i/4}$. Hence we have to prove only that $U \in \mathfrak{B}^2_{\bar{\beta}}(S_{i,\delta_i/4})$.

Let

$$\varphi_0 \in C^\infty(\mathbb{R}^+),$$

$$\varphi_0(r) = 1 \quad \text{for } 0 \leqq r \leqq \tfrac{1}{2},$$

$$\varphi_0(r) = 0 \quad \text{for } x \geqq 1,$$

$$\varphi_{\delta_i}(r) = \varphi_0(r/2\delta_i) = \varphi(r).$$

Denote $v = \varphi U$. Then v can be understood to be defined on the infinite sector $Q^{(i)}_{w_i} = \{(r_i, \theta_i) | 0 < r_i < \infty, 0 < \theta_i < \omega_i\}$ when extended by zero outside of S_{i,δ_i} and we have $v \in H^1(Q^{(i)}_{\omega_i})$. □

Now we prove that $v \in \mathfrak{B}^2_\beta(S_{i,\delta_i/2})$ as in [3].

Remark 4.3. We have assumed that either $g \in \mathfrak{B}^1_{\hat{\beta}}(\Gamma_1)$, $0 < \hat{\beta} < \frac{1}{2}$ or $g \in \mathfrak{B}^2_{\hat{\beta}}(\Gamma_1)$, $\frac{1}{2} < \hat{\beta} < 1$. Obviously Theorem 4.2 is correct if $g \in \mathfrak{B}^1_{\hat{\beta}}(\Gamma_1)$ only in the neighborhood of A_1 and $g \in \mathfrak{B}^2_{\hat{\beta}}(\Gamma_1)$ in the neighborhood of A_2. Theorem 4.1 leads easily to the next theorem.

THEOREM 4.3. *Let Ω be a straight polygon with the edges Γ_i, $i = 1, \cdots, M$ and let*

$$g \in \mathfrak{B}^1_{\bar{\beta}_i}(\Gamma_i), \quad \hat{\beta}_i = (\hat{\beta}_{i,1}, \hat{\beta}_{i,2}), \quad 0 < \hat{\beta}_{i,1}, \hat{\beta}_{i,2} < \tfrac{1}{2},$$
$$\bar{\beta}_{i,1} = \hat{\beta}_{i,1} + \tfrac{1}{2}, \quad \bar{\beta}_{i,2} = \hat{\beta}_{i,2} + \tfrac{1}{2}, \quad i \in Q \subset \{1, \cdots, M\}$$

or

$$g \in \mathfrak{B}^2_{\bar{\beta}_i}(\Gamma_i), \quad \hat{\beta}_i = (\hat{\beta}_{i,1}, \hat{\beta}_{i,2}), \quad \tfrac{1}{2} < \hat{\beta}_{i,1}, \hat{\beta}_{i,2} < 1,$$
$$\bar{\beta}_{i,1} = \hat{\beta}_{i,1} - \tfrac{1}{2}, \quad \bar{\beta}_{i,2} = \hat{\beta}_{i,2} - \tfrac{1}{2}, \quad i \in Q \subset \{1, \cdots, M\}.$$

Further, let g be continuous on $\gamma = \bigcup_{i \in Q} \bar{\Gamma}_i$. Then $g \in \mathfrak{B}^{3/2}_{\bar{\beta}}(\gamma)$ where $\bar{\beta}_i = \max(\bar{\beta}_{i-1,2}, \bar{\beta}_{i,1})$, for $A_i \in \gamma$ (if $i - 1 \notin Q$ or $i \notin Q$ then we define $\bar{\beta}_{i-1,2} = 0$, respectively, $\bar{\beta}_{i,1} = 0$) and $0 < \bar{\beta}_i < 1$ arbitrarily for $A_i \notin \gamma$.

Proof. Because g is continuous on γ we can construct a polynomial P on Ω such that $g - P = 0$ at A_i. Then we can apply Theorem 4.2. □

Remark 4.4. It is obvious how the theorem may be modified when $g \in \mathfrak{B}^1_\beta(\Gamma_i)$, respectively, $g \in \mathfrak{B}^2_\beta(\Gamma_i)$ in the neighborhood of A_i only. See also Remark 4.3.

Remark 4.5. Theorems 4.1 and 4.3 are complementary, which is analogous to the theorems of trace and extension in usual Sobolev spaces on smooth domain. Namely, if $g \in \mathfrak{B}^1_{\hat{\beta}_i}(\Gamma_i)$, $0 < \hat{\beta}_{i,j} < \frac{1}{2}$ (respectively, $g \in \mathfrak{B}^2_{\hat{\beta}_i}(\Gamma)$, $\frac{1}{2} < \hat{\beta}_{i,j} < 1$) $j = 1, 2$, then we have an extension by function $G \in \mathfrak{B}^2_\beta(\Omega)$, $\beta_i = \hat{\beta}_{i,1} + \frac{1}{2}$, $\beta_{i+1} = \hat{\beta}_{i,2} + \frac{1}{2}$ (respectively, $\beta_i = \hat{\beta}_{i,1} - \frac{1}{2}$, $\beta_{i+1} = \hat{\beta}_{i,2} - \frac{1}{2}$), and if $G \in \mathfrak{B}^2_\beta(\Omega)$ then $G|_{\Gamma_i} = g \in \mathfrak{B}^1_{\hat{\beta}_i + \varepsilon}(\Gamma_i)$, $\hat{\beta}_{i,1} = \beta_i - \frac{1}{2}$, $\hat{\beta}_{i,2} = \beta_{i+1} - \frac{1}{2}$ for $\frac{1}{2} < \beta_i, \beta_{i+1} < 1$ (respectively, $g \in \mathfrak{B}^2_{\hat{\beta}_i + \varepsilon}(\Gamma_i)$, $\hat{\beta}_{i,1} = \beta_i + \frac{1}{2}$, $\hat{\beta}_{i,2} = \beta_{i+1} + \frac{1}{2}$ for $0 < \beta_i, \beta_{i+1} < \frac{1}{2}$), $\varepsilon > 0$ arbitrary.

THEOREM 4.4. *Let Ω be a straight polygon with the edges Γ_i, $i=1,\cdots,M$, and let $g\in\mathfrak{B}^0_{\hat\beta}(\Gamma_1)$, $0<\hat\beta_i<\frac12$, $i=1,2$, $\beta_i=\hat\beta_i+\frac12$, $i=1,2$ (respectively, $g\in\mathfrak{B}^1_{\hat\beta}(\Gamma_1)$, $\frac12<\hat\beta_i<1$, $\beta_i=\hat\beta_i-\frac12$, $i=1,2$). Then there is u such that we have the following:*

(i) *$u\in\mathfrak{B}^1_\beta(\Omega)$ with $0<\beta_j<1$, $j=3,\cdots,M$ arbitrary.*

(ii) *$u|_{\Gamma_i}=g$ and $u|_{\Gamma_j}=0$, $j=2,\cdots,M$.*

Proof. By Lemma 4.14, $\tilde g=g\hat\Phi\in\mathfrak{B}^1_{\hat\beta}(\Gamma_1)$, respectively, $\mathfrak{B}^2_{\hat\beta}(\Gamma_1)$ and $\tilde g(A_i)=0$, $i=2,3$, and hence by Theorem 4.2 there is $v\in\mathfrak{B}^2_\beta(\Omega)$ such that $v=\tilde g$ on Γ_1 and $v=0$ on Γ_j, $j=2,\cdots,M$. By Lemma 4.15 the function $v\Phi^{-1}$ has the desired properties. □

Theorem 4.4 leads immediately to Theorem 4.5.

THEOREM 4.5. *Let Ω be a straight polygon with the edges Γ_i, $i=1,\cdots,M$, and let*

$$g\in\mathfrak{B}^0_{\hat\beta_i}(\Gamma_i),\quad \hat\beta_i=(\hat\beta_{i,1},\hat\beta_{i,2}),\quad 0<\hat\beta_{i,1},\hat\beta_{i,2}<\tfrac12,$$

$$\bar\beta_{i,1}=\hat\beta_{i,1}+\tfrac12,\quad \bar\beta_{i,2}=\hat\beta_{i,2}+\tfrac12,\quad i\in Q\subset\{1,\cdots,M\}$$

or

$$g\in\mathfrak{B}^1_{\hat\beta_i}(\Gamma_i),\quad \hat\beta_i=(\hat\beta_{i,1},\hat\beta_{i,2}),\quad \tfrac12<\hat\beta_{i,1},\hat\beta_{i,2}<1,$$

$$\bar\beta_{i,1}=\hat\beta_{i,1}-\tfrac12,\quad \bar\beta_{i,2}=\hat\beta_{i,2}-\tfrac12,\quad i\in Q\subset\{1,\cdots,M\}.$$

Let $\gamma=\bigcup_{i\in Q}\bar\Gamma_i$. Then $g\in\mathfrak{B}^{1/2}_\beta(\gamma)$ where $\bar\beta_i=\max(\bar\beta_{i-1,2},\bar\beta_{i,1})$, $A_i\in\gamma$ (if $i-1\notin Q$ or $i\notin Q$ when we define $\hat\beta_{i-1,2}=0$, respectively, $\bar\beta_{i,1}=0$) and $0<\bar\beta_i<1$ arbitrarily for $A_i\notin\gamma$.

Remark 4.6. It is obvious how Theorem 4.4 has to be modified when $g\in\mathfrak{B}^1_{\hat\beta}(\Gamma_i)$, respectively, $g\in\mathfrak{B}^2_{\hat\beta}(\Gamma_i)$ in the neighborhood of A_i only. See Remark 4.3.

Theorems 4.3 and 4.5 give the characterization of the boundary conditions that guarantees that the solution of an elliptic partial differential equation of second order with analytic coefficients on a domain Ω with piecewise analytic boundary belongs to $\mathfrak{B}^2_\beta(\Omega)$ or $\mathfrak{C}^2_\beta(\Omega)$ (see Theorems 3.2 and 3.3).

In the concrete cases these conditions are usually very easy to check. Let us state a useful lemma that characterizes the space $\mathfrak{B}^1_{\hat\beta}(I)$ (respectively, $\mathfrak{B}^2_{\hat\beta}(I)$).

LEMMA 4.16. *Let*

$$\Omega_\alpha=\{z=x+iy\,|\,x\in I,\ |y|\leqq\alpha\hat\Phi(x),\ \alpha>0\}$$

and $G(z)$ be a holomorphic function on Ω_α such that for $\nu=(\nu_1,\nu_2)$

$$|G(z)|\leqq C\hat\Phi_\nu(\operatorname{Re} z).$$

Let $g(x)=\operatorname{Re} G(z)|_I$ or $\operatorname{Im} G(z)|_I$. Then for $\nu_i>-\frac12+(j-1)$, $\hat\beta_i+\nu_i>\frac12+(j-1)$, $0<\hat\beta_i<1$, $i=1,2$, $j=0,1,2$

$$g(x)\in\mathfrak{B}^j_{\hat\beta}(I).$$

Proof. By the Cauchy formula we have for $k>0$

$$|g^{(k)}(x)|\leqq C\hat\Phi_\nu(x)(\hat\Phi(x))^{-k}k!\alpha^{-k}.$$

Hence,

$$\int_{-1}^1\Phi^2_{k-1+\hat\beta}|g^{(k)}(x)|^2\,dx\leqq(Ck!\alpha^{-k})^2\int_{-1}^1\Phi^2_{\nu+\hat\beta-1}\,dx\leqq(C_1d^kk!)^2$$

provided that $\nu_i+\hat\beta_i>\frac12$. Furthermore, for $k=0$

$$|g(x)|\leqq C\Phi_\nu(x),$$

and hence for $\nu_i > -\frac{1}{2}$, $g \in H^0(I)$. The lemma is proved for $j=1$. The proof of the case $j=0$ is analogous. Let us consider now the case $j=2$. We see that for $\nu_i + \hat{\beta}_i > \frac{3}{2}$ and $k \geqq 2$

$$\int_{-1}^{1} \Phi^2_{k-2+\hat{\beta}} |g^{(k)}(x)|^2 \, dx \leqq (Ck!\alpha^{-k})^2 \int_{-1}^{1} \Phi^2_{\nu-k+k-2+\hat{\beta}} \, dx \leqq (C_1 d^k k!)^2.$$

Furthermore, if $\nu_i > \frac{1}{2}$, then also $g \in H^1(I)$. □

Instead of $|G(z)| \leqq C\hat{\Phi}_\nu(\operatorname{Re} z)$ we can assume that $|G(z) - P(z)| \leqq C\hat{\Phi}_\nu(\operatorname{Re} z)$ where $P(z)$ is a polynomial.

Lemma 4.16 is very useful in practice. For example, if g is analytic on $\bar{\Gamma}_i$ then $g(x)$ can be extended into some neighborhood of $\bar{\Gamma}_i$ and therefore $g \in \mathfrak{B}^1_\beta(I)$. Lemma 4.16 characterizes very well the structure of the spaces $\mathfrak{B}^1_\beta(I)$ (respectively, $\mathfrak{B}^2_\beta(I)$).

LEMMA 4.17. *Let $g \in \mathfrak{B}^1_{\hat{\beta}}(I)$, $0 < \hat{\beta}_i < \frac{1}{2}$. Then there exists $\alpha > 0$ such that g can be analytically extended onto Ω_α and*

$$\left| G(z) - g(-1)\frac{(1-x)}{2} - g(1)\frac{(x+1)}{2} \right| \leqq C\hat{\Phi}_{1/2-\hat{\beta}}(\operatorname{Re} x)$$

($g \in C^0(\bar{I})$ by Theorem 3.1).

Proof. Since $g \in \mathfrak{B}^1_{\hat{\beta}}(I)$ we have by Lemma 4.3 for $k \geqq 1$

$$|g^{(k)}(x)| \leqq C(\Phi_{k-1/2+\hat{\beta}}(x))^{-1} d^k k!.$$

Hence the series

$$g'(x) = \sum_{k=0}^{\infty} g^{(k+1)}(x_0)(x-x_0)^k \frac{1}{k!}, \qquad x_0 \in I$$

is absolutely convergent for $|x - x_0| \leqq \frac{1}{2}(\Phi(x_0)/d)$, and hence also

$$G'(z) = \sum_{k=0}^{\infty} g^{(k+1)}(x_0)(z-x_0)^k \frac{1}{k!}$$

converges for $|z - x_0| \leqq \frac{1}{2}(\Phi(x_0)/d)$ and $|G'(z)| \leqq C\hat{\Phi}^{-1}_{\hat{\beta}+1/2}(x_0)$, $x_0 = \operatorname{Re}(z)$, and C is independent of x_0, which yields the lemma. □

So far we have assumed that Ω is a straight polygon. We did not exclude the case where the internal angle is 2π; i.e., we did not exclude the slit domain. Let us now consider the curvilinear polygon and assume that it is a Lipschitzian domain. Let us prove first Lemma 4.18.

LEMMA 4.18. *Let* $\Omega = \{x_1, x_2 | -1 < x_1 < 1, 0 < x_2 < h(x_1), h(x_1) > \alpha(x_1+1), h(-1) = 0, \alpha > 0\}$. *Assume that $\psi(x_1, x_2)$ is an analytic function on* $S = \{x_1, x_2 | (x_1+1)^2 + x_2^2 \leqq 4\}$ *such that we have the following*:

(i) $\psi(x_1, h(x_1)) = 0$;

(ii) $\partial\psi/\partial x_1(x_1, 0) > \alpha > 0$, $-1 \leqq x_1 \leqq 1$.

Define

$$\Gamma_1 = \{x_1, x_2 | -1 < x_1 < 1, x_2 = 0\},$$

$$\Gamma_2 = \{x_1, x_2 | -1 < x_1 < 1, x_2 = h(x_1)\},$$

and let $T = \Omega \cap S_1$ where $S_1 = \{r, \theta | 0 < \theta < 2\pi, 0 < r < 1\}$ where (r, θ) are polar coordinates with respect to $(-1, 0)$ and $T^* = S_1 - T$. Let $g_1 \in \mathfrak{B}^1_{\hat{\beta}}(\Gamma_1)$, $0 < \hat{\beta}_i < \frac{1}{2}$, $\hat{\beta}_1 = \hat{\beta}_2$ (respectively, $g_2 = \mathfrak{B}^2_{\hat{\beta}}(\Gamma_1)$, $\frac{1}{2} < \hat{\beta}_i < 1$, $\hat{\beta}_1 = \hat{\beta}_2$), $g_i(-1) = 0$, $i = 1, 2$ and $\Phi = r$.

Then there exists

$$V_1 \in \mathfrak{B}^2_{\bar{\beta}}(T), \quad V_1^* \in \mathfrak{B}^2_{\bar{\beta}}(T^*), \quad \bar{\beta} = \hat{\beta} + \tfrac{1}{2}$$

(respectively, $V_2 \in \mathfrak{B}^2_{\bar\beta}(T)$, $V_2^* \in \mathfrak{B}^2_{\bar\beta}(T^*)$, $\bar\beta = \hat\beta - \frac{1}{2}$*) such that* $V_i = g_i$ *and* $V_i^* = g_i$ *on* $\Gamma_1 \cap \bar T$ *and* V_i, $V_i^* = 0$ *on* $\Gamma_2 \cap \bar T$.

Proof. Let $\varphi(r, \theta) = \psi(r, \theta)/r$. Then $\varphi(r, 0) = \varphi(x_1)$ is analytic on $\bar\Gamma_1$ and $\varphi(x_1) > \bar\alpha > 0$; hence $\varphi^{-1}(x_1)$ is analytic on $\bar\Gamma_1$ too. In addition, $\varphi = 0$ on Γ_2. Furthermore, $|D^\alpha \varphi(x_1, x_2)| \leqq C|\alpha|!\Phi^{-|\alpha|} d^{|\alpha|}$ by Cauchy's theorem on the theory of two complex variables. Define $\tilde g_1 = g_1 \varphi^{-1}(x_1)$. Then $\tilde g_1 \in \mathfrak{B}^1_{\hat\beta}(\Gamma_1)$ and by Lemma 4.11 there exists U_1 on S_1 such that $U_1 \in \mathfrak{B}^2_{\bar\beta}(S_1)$, $\bar\beta = \hat\beta + \frac{1}{2}$ and $U_1|_{\Gamma_1} = \tilde g_1$. Now define $V_1 = U_1 \varphi$. Using Lemma 4.13, we conclude that $V_1 \in \mathfrak{B}^2_{\bar\beta}(T)$ (respectively, $\mathfrak{B}^2_{\bar\beta}(T^*)$), $V_1|_{\Gamma_1} = g_1$ and $V_1|_{\Gamma_2} = 0$. The proof that V_2 has desired properties is analogous. □

LEMMA 4.19. *Let* $\Omega = \{x_1, x_2 | -1 < x_1 < 1, h_1(x_1) < x_2 < h_2(x_1), h_1(x_1) < -\alpha(x_1 + 1), h_2(x_1) > \alpha(x_1 + 1), \alpha > 0, h_i(-1) = 0, h_i(x_1)$ *analytic functions on* $\bar I, i = 1, 2\}$ *and*

$$\Gamma_i = \{x_1, x_2 | -1 < x_1 < 1, x_2 = h_i(x_1)\},$$

$$\Omega_\eta = \Omega \cap S_\eta, \quad S_\eta = \{r, \theta | 0 < \theta \leqq 2\pi, 0 < r < \eta, \eta > 0\}, \quad \Omega_\eta^* = S_\eta - \Omega_\eta,$$

where (r, θ) *are polar coordinates with the origin at* $(-1, 0)$. *Let* $g_1 \in \mathfrak{B}^1_{\hat\beta}(\Gamma_1)$, $0 < \hat\beta < \frac{1}{2}$ *(respectively,* $g_2 \in \mathfrak{B}^2_{\hat\beta}(\Gamma_1)$, $\frac{1}{2} < \hat\beta_i < 1$*),* $\beta_1 = \beta_2$, $g_i(-1) = 0$ *and let* $\Phi = r$. *Then there exist* $\eta > 0$ *and* $V_1 \in \mathfrak{B}^2_{\bar\beta}(\Omega_\eta)$, $V_1^* \in \mathfrak{B}^2_{\bar\beta}(\Omega_\eta^*)$, $\bar\beta = \hat\beta + \frac{1}{2} + \varepsilon$, $\varepsilon > 0$ *arbitrary (respectively,* $V_2 \in \mathfrak{B}^2_{\bar\beta}(\Omega_\eta)$, $V_2 \in \mathfrak{B}^2_{\bar\beta}(\Omega_\eta^*)$, $\bar\beta = \hat\beta - \frac{1}{2} + \varepsilon$*) such that* $V_i|_{\Gamma_1 \cap \bar\Omega_\eta} = g_i$ *and* $V_i|_{\Gamma_2 \cap \bar\Omega_\eta} = 0$.

Proof. Because $h_1(x_1)$ is analytic on $\bar I$ it can be analytically extended onto $\tilde I_\delta = \{-1 - \delta < x_1 < 1 + \delta\}$. Then the mapping $M: (x_1, x_2) \to (y_1, y_2)$, $y_1 = x_1$, $y_2 = x_2 - h_1(x_1)$ is analytic on Ω_η, $\eta = \delta/2$ and $M(\Omega_\eta) = \tilde\Omega_\eta$. For η_1 sufficiently small we have $\partial\tilde\Omega_\eta \cap S_{\eta_1} = \Gamma_1^* \cup \Gamma_2^*$ where

$$\Gamma_1^* = \{y_1, y_2 | -1 < y_1 < -1 + \eta, y_2 = 0\},$$

$$\Gamma_2^* = \{y_1, y_2 | -1 < y_1 < -1 + \bar\eta_1, y_2 = h_2^*(y_1) = h_2(y_1) - h_1(y_1)\},$$

and $h_2(y_1) > \alpha_1(y_1 + 1)$. In addition, it is easy to see that $|J|$, $|J^{-1}| < \mu < \alpha$, where J is the Jacobian of the mapping M. Because $h_2^*(y_1)$ is analytic on $-1 \leqq y_1 \leqq -1 + \bar\eta_1$ we define $\psi(y_1, y_2) = -y_2 + h_2^*(y_1)$ and $\psi(y_1, y_2)$ has the properties in Lemma 4.18. Now using Corollaries 4.4, 4.5, $g_1 \in \mathfrak{C}^1_{\hat\beta}(\Gamma_1)$, $g_2 \in \mathfrak{C}^2_{\hat\beta}(\Gamma_1)$, and hence using Lemma 4.6, $g_1(M^{-1}(y))|_{y_2=0} \in \mathfrak{C}^1_{\hat\beta}(\Gamma_1^*)$, $g_2(M^{-1}(y))|_{y_2=0} \in \mathfrak{C}^2_{\hat\beta}(\Gamma_1^*)$. Using Lemmas 4.8 and 4.18, we obtain functions V_1 and V_1^* (respectively, V_2 and V_2^*) on $\tilde\Omega_\eta \cap S_{\eta_2}$ (respectively, $\tilde\Omega_\eta^* \cap S_{\eta_2}$), which belong to $\mathfrak{B}^2_{\bar\beta+\varepsilon/2}(\Omega_\eta \cap S_{\eta_2})$ (respectively, $\mathfrak{B}^2_{\bar\beta+\varepsilon/2}(\tilde\Omega_\eta^* \cap S_{\eta_2})$). Now when we use Lemmas 4.7 and 2.3, our lemma follows. □

The lemma leads to the following theorem.

THEOREM 4.6. *Theorems* 4.3 *and* 4.5 *hold also for a Lipschitzian curvilinear polygon when* $\bar\beta_i$ *are replaced by* $\bar\beta_i + \varepsilon$, $\varepsilon > 0$ *arbitrary.*

Proof. Because the edges are analytic curves and g are analytic on Γ_i (but not on $\bar\Gamma_i$) we show similarly (as in the proof of Theorem 4.1) that the solution u of the Laplace equation belongs to $\mathfrak{B}^2_{\bar\beta+\varepsilon}(\Omega)$. This can be done identically as in the proofs of Theorems 3.3 and 3.4 of [6], showing that $u \in \mathfrak{C}^2_{\bar\beta+\varepsilon}(\Omega)$. □

Remark 4.7. Comparing the respective theorems for straight and curvilinear polygons, we see that in the latter case we lose slightly in the regularity. It is not known whether this loss can be removed.

REFERENCES

[1] I. BABUŠKA, *The p and h-p version of the finite element method; the state of the art*, in "Finite Elements, Theory and Applications," D. L. Dwyer, M. Y. Hussaini, and R. G. Voigt, eds., Springer-Verlag, Berlin, New York, 1988, pp. 199-239.

[2] I. BABUŠKA AND M. R. DORR, *Error estimates for the combined h and p version of the finite element method*, Numer. Math., 37 (1981), pp. 257-277.

[3] I. BABUŠKA AND B. Q. GUO, *Regularity of the solution of elliptic problems with piecewise analytic data. Part* 1: *Boundary value problems for linear elliptic equation of the second order*, SIAM J. Math. Anal., 19 (1988), pp. 172-203.

[4] ———, *The h-p version of finite element method with curved boundary*, SIAM J. Numer. Anal., 25 (1988), pp. 837-861.

[5a] B. Q. GUO AND I. BABUŠKA, *The h-p version of finite element method, part* 1: *the basic approximation results*, Comp. Mech., 1 (1986), pp. 21-41.

[5b] ———, *The h-p version of finite element method, part* 2: *general results and applications*, Comp. Mech., 1 (1986), pp. 203-220.

[6] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series and Products*, Academic Press, New York, 1980.

[7] G. H. HARDY, J. E. LITTLEWOOD, AND G. POLYA, *Inequality*, Second edition, Cambridge University Press, London, 1952.

[8] C. B. MORREY, *Multiple Integrals in Calculus of Variations*, Springer-Verlag, Berlin, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 782-788, July 1989

002

DECAY ESTIMATES FOR A CLASS OF NONLINEAR BOUNDARY VALUE PROBLEMS IN TWO DIMENSIONS*

C. O. HORGAN† AND L. E. PAYNE‡

Abstract. This paper is concerned with second-order quasilinear partial differential equations in two independent variables of the form $\operatorname{div}[\rho(\mathbf{x}, u, \operatorname{grad} u)\operatorname{grad} u] = 0$. Previous work of the authors, establishing exponential decay estimates for Dirichlet problems on a semi-infinite strip subject to nonzero data on the finite end, is extended to include regions of arbitrary shape, and, in the case of unbounded regions, the a priori assumption that solutions must decay to zero as $|\mathbf{x}| \to \infty$ is removed. The results have application to Saint-Venant principles for nonlinear elasticity as well as to theorems of Phragmén-Lindelöf type.

Key words. second-order quasilinear partial differential equations, two independent variables, decay estimates, Dirichlet problems, Saint-Venant principles, Phragmén-Lindelöf type theorems

AMS(MOS) subject classifications. 35B40, 35J60, 73C10

1. Introduction. In a recent paper [3] the authors derived exponential decay estimates for a certain class of nonlinear boundary-value problems in the plane. Specifically, they treated a class of second-order quasilinear equations (not necessarily elliptic) defined in a semi-infinite strip with nonzero boundary data at the finite end. Assuming boundedness of certain functionals on the strip, the authors derived exponential decay estimates in various norms. In a subsequent paper [4] these results were extended to $\mathbb{R}^3$.

In this paper we remove the requirement that the region be a strip domain (i.e., of constant cross section) and in the case of unbounded regions we eliminate the a priori assumption that solutions must decay to zero uniformly as $|\mathbf{x}| \to \infty$, where $|\mathbf{x}| = (x_1^2 + x_2^2)^{1/2}$.

The spatial decay estimates of concern in [3], [4] and in the present paper are of interest in connection with studies on Saint-Venant's principle in elasticity theory. (See [2] for a review of results on Saint-Venant's principle.) The results may also be viewed as giving rise to theorems of Phragmén-Lindelöf type (see, e.g., the references cited in [3]).

Let Ω be a bounded (or unbounded) simply-connected region with Lipschitz boundary $\partial\Omega$, lying in the half plane $x_2 > 0$, and let $u(x_1, x_2)$ be a classical solution of the equation

$$[\rho(\mathbf{x}, u, \operatorname{grad} u)u_{,j}]_{,j} = 0 \qquad (j = 1, 2) \tag{1.1}$$

in Ω. Here we have adopted the convention of summing over repeated indices and a comma denotes partial differentiation. The boundary $\partial\Omega$ is composed of a small arc Σ on which no data are prescribed and the remainder $\partial\Omega/\Sigma$ on which u is assumed to vanish. When Ω is unbounded it is assumed that no data are prescribed on Σ and that u vanishes on the finite portion of the remainder of $\partial\Omega$. For the moment no a priori assumption is made about the behavior of u as $|\mathbf{x}| \to \infty$ in Ω, where $|\mathbf{x}| = (x_1^2 + x_2^2)^{1/2}$.

* Received by the editors August 17, 1987, accepted for publication (in revised form) June 1, 1988.

† Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia 22903. The research of this author was supported by National Science Foundation grant MSM-85-12825 and by the Center for Advanced Studies, University of Virginia, Charlottesville, Virginia 22903.

‡ Department of Mathematics, Cornell University, Ithaca, New York 14853. The research of this author was supported by National Science Foundation grant DMS-86-00250.

Although we assume existence of a classical solution throughout, it will be clear that the same decay results would hold for appropriately defined weak solutions.

Consider then the solution u of (1.1) subject to the condition

$$u = 0 \quad \text{on } \partial\Omega/\Sigma. \tag{1.2}$$

We wish to derive a bound for a suitable norm of the solution that decays as some function of the distance from Σ. The assumptions on the form of ρ are as follows. We have either Case 1 or Case 2 as follows.

Case 1.

$$0 < m_1 \leqq \rho \leqq M_1 + K_1 \rho q^2. \tag{1.3}$$

Case 2.

$$0 < m_2 \leqq \rho^{-1} \leqq M_2 + K_2 \rho q^2 \quad \text{in } \Omega. \tag{1.4}$$

Here we have used the notation $q^2 = |\text{grad } u|^2$. These are the same assumptions made in [3]. It will be clear that M_1 and M_2 could be allowed to depend on x_1 and x_2, but in this paper we assume that the m_i and M_i are positive constants and that the K_i are nonnegative constants. We recall the discussion in [3] concerning assumptions (1.3), (1.4). If, for instance, ρ were a bounded function of its arguments, then K_1 in (1.3) could be taken to be zero. Roughly speaking, the first term on the right in (1.3) provides a bound on ρ as $q \to 0$, while the second term gives a bounding function for ρ as $q \to \infty$. A function ρ for which (1.3) holds is $\rho = 1 + q^2$, in which case we may take $m_1 = 1$, $M_1 = 1$, and $K_1 = 1$. If $\rho = (1+q^2)^{-1/2}$, in which case (1.1) is the minimal surface equation, then (1.4) is satisfied with $m_2 = 1$, $M_2 = 1$, and $K_2 = 1$.

Let us now introduce a family of curves $f(x_1, x_2) = \alpha$, $\alpha > \alpha_0$. For $\alpha_0 < \alpha < \alpha_1$ the curve $f = \alpha$ is assumed to intersect Ω and form a region Ω_α defined as

$$\Omega_\alpha = \Omega \cap \{x_1, x_2; f(x_1, x_2) > \alpha\}. \tag{1.5}$$

It is further assumed that through each point of Ω one and only one curve of the family passes and that the following properties hold:

(i) $\beta \leqq \delta \Rightarrow \Omega_\beta \supset \Omega_\delta$;

(ii) $\partial\Omega_{\alpha_0}$ contains no points of Σ;

(iii) $|\text{grad } f| \leqq \gamma$ in Ω_{α_0}.

We will also investigate the case in which Ω is unbounded and α_1 is infinity. In the previous paper [3] the region Ω was assumed to be the semi-infinite strip $\{x_1 > 0, 0 < x_2 < h\}$, and $f(x_1, x_2)$ was chosen to be the coordinate x_1. For certain geometries this same choice of f can be made for the class of problems considered in this paper.

2. Decay estimates in the first case. The method of proof will differ somewhat from that used in [3]. Furthermore, for the case of unbounded Ω we make no a priori assumption on the decay of u as $|\mathbf{x}| \to \infty$ in Ω. The method itself will provide an appropriate alternative.

Let us first introduce some additional notation, i.e.,

$$\begin{aligned} S_\alpha &:= \{f(x_1, x_2) = \alpha\} \cap \Omega, \\ l(\alpha) &= \text{length of } S_\alpha, \\ B(\alpha) &= \int_{S_\alpha} \rho |\text{grad } f| \, ds, \\ \mathscr{B}(\alpha) &= \int_{S_\alpha} \rho^{-1} |\text{grad } f| \, ds. \end{aligned} \tag{2.1}$$

We wish to consider the behavior of the functional

$$F(\alpha)=\int_{S_\alpha}\frac{\rho u u_{,i} f_{,i}}{|\operatorname{grad} f|}\,ds, \qquad \alpha_0<\alpha<\alpha_1 \tag{2.2}$$

as a function of α. It is understood throughout that $\rho \equiv \rho(\mathbf{x}, u, \operatorname{grad} u)$.

Now, $F(\alpha)$ may be rewritten as either

$$F(\alpha)=-\int_{\Omega_\alpha}\rho q^2\,dx \quad \text{if } \Omega \text{ is bounded,} \tag{2.3}$$

or

$$F(\alpha)=F(\alpha_0)+\int_{\alpha_0}^{\alpha}\int_{S_\eta}\frac{\rho q^2}{|\operatorname{grad} f|}\,ds\,d\eta \quad \text{if } \Omega \text{ is unbounded.} \tag{2.4}$$

We could, of course, use the second representation in either case. Both representations follow directly from the divergence theorem. The weighted Dirichlet integrals (2.3), (2.4) provide a measure for the "energy" associated with solutions of (1.1), (1.2). It follows that

$$F'(\alpha)=\int_{S_\alpha}\frac{\rho q^2}{|\operatorname{grad} f|}\,ds, \tag{2.5}$$

where the prime denotes differentiation with respect to α. Our first objective is to derive an explicit inequality of the form

$$|F(\alpha)|\leqq k^{-1}(\alpha)F'(\alpha). \tag{2.6}$$

This will lead to the following two first-order differential inequalities:

$$F'(\alpha)\geqq k(\alpha)F(\alpha), \tag{2.7}$$

$$F'(\alpha)\geqq -k(\alpha)F(\alpha). \tag{2.8}$$

One of these inequalities will be superfluous if Ω is bounded, but it will permit us to derive an alternative result if Ω is unbounded. We now proceed to establish (2.6).

By Schwarz's inequality,

$$F^2(\alpha)\leqq\int_{S_\alpha}\frac{\rho(\partial u/\partial n)^2}{|\operatorname{grad} f|}\,ds\int_{S_\alpha}\rho u^2|\operatorname{grad} f|\,ds. \tag{2.9}$$

Regarding ρ as a function of the variables (x_1, x_2) and introducing the variable

$$\sigma=\int_0^s \rho|\operatorname{grad} f|\,ds, \tag{2.10}$$

we have

$$\int_{S_\alpha}\rho u^2|\operatorname{grad} f|\,ds=\int_0^{B(\alpha)}u^2\,d\sigma\leqq\frac{[B(\alpha)]^2}{\pi^2}\int_0^{B(\alpha)}\left(\frac{\partial u}{\partial\sigma}\right)^2 d\sigma, \tag{2.11}$$

where $B(\alpha)$ is as defined in (2.1). The boundary condition (1.2) and a well-known inequality of Wirtinger type (see e.g., [1, p. 185]) have been used in obtaining the last inequality. Reverting to the original coordinates, we have

$$\begin{aligned}\int_{S_\alpha}\rho u^2|\operatorname{grad} f|\,ds&\leqq\frac{[B(\alpha)]^2}{\pi^2}\int_{S_\alpha}\rho^{-1}\left(\frac{\partial u}{\partial s}\right)^2|\operatorname{grad} f|^{-1}\,ds\\&\leqq\frac{[B(\alpha)]^2}{m_1^2\pi^2}\int_{S_\alpha}\frac{\rho}{|\operatorname{grad} f|}\left(\frac{\partial u}{\partial s}\right)^2 ds,\end{aligned} \tag{2.12}$$

where the left-hand inequality of (1.3) has been used in the last step. Insertion of the above into (2.9) leads to

$$
\begin{aligned}
|F(\alpha)| &\leqq \frac{B(\alpha)}{m_1\pi}\left\{\int_{S_\alpha}\frac{\rho}{|\operatorname{grad} f|}\left(\frac{\partial u}{\partial n}\right)^2 ds \int_{S_\alpha}\frac{\rho}{|\operatorname{grad} f|}\left(\frac{\partial u}{\partial s}\right)^2 ds\right\}^{1/2} \\
&\leqq \frac{B(\alpha)}{2m_1\pi}F'(\alpha).
\end{aligned} \tag{2.13}
$$

We have thus established (2.6) with $k(\alpha)=2m_1\pi/B(\alpha)$.

An integration of (2.7) with the above value for $k(\alpha)$ leads to

$$
F(\alpha)\geqq F(\tilde{\alpha})\exp\left\{2m_1\pi\int_{\tilde{\alpha}}^{\alpha}[B(\eta)]^{-1}\,d\eta\right\} \tag{2.14}
$$

for $\alpha\geqq\tilde{\alpha}>\alpha_0$. This inequality shows that if $F(\alpha)>0$ for any α (say $\tilde{\alpha}$), then $F(\alpha)>0$ for $\alpha>\tilde{\alpha}$, and hence if Ω is bounded then $F(\alpha_1)$ cannot be zero. Thus in this case $F(\alpha)<0$, for all α. It is, of course, obvious from (2.3) that, for bounded Ω, $F(\alpha)$ can never be positive. Thus for Ω bounded we have from (2.8)

$$
-F(\alpha)\leqq -F(\alpha_0)\exp\left\{-2m_1\pi\int_{\alpha_0}^{\alpha}[B(\eta)]^{-1}\,d\eta\right\}, \tag{2.15}
$$

which we write as

$$
\int_{\Omega_\alpha}\rho q^2\,dx\leqq\int_{\Omega_{\alpha_0}}\rho q^2\,dx\exp\left\{-2m_1\pi\int_{\alpha_0}^{\alpha}[B(\eta)]^{-1}\,d\eta\right\}. \tag{2.16}
$$

The exponential will now be bounded in a manner somewhat different from that used in [3]. We note that

$$
\begin{aligned}
\int_{\alpha_0}^{\alpha}[B(\eta)]^{-1}\,d\eta &\geqq \int_{\alpha_0}^{\alpha}d\eta\Big/\left[\int_{S_\eta}\{M_1+K_1\rho q^2\}|\operatorname{grad} f|\,ds\right] \\
&\geqq \frac{1}{M_1\gamma}\int_{\alpha_0}^{\alpha}d\eta\Big/ l(\eta)\left\{1+\frac{K_1\gamma}{M_1 l(\eta)}\int_{S_\eta}\frac{\rho q^2}{|\operatorname{grad} f|}\,ds\right\},
\end{aligned} \tag{2.17}
$$

where we recall from (2.1) that $l(\eta)$ is the length of S_η. Thus

$$
\begin{aligned}
\int_{\alpha_0}^{\alpha}[B(\eta)]^{-1}\,d\eta &\geqq \frac{1}{M_1\gamma}\int_{\alpha_0}^{\alpha}\frac{d\eta}{l(\eta)}-\frac{K_1}{M_1^2}\int_{\alpha_0}^{\alpha}\left(\int_{S_\eta}\frac{\rho q^2}{l^2(\eta)|\operatorname{grad} f|}\,ds\right)d\eta \\
&\geqq \frac{1}{M_1\gamma}\int_{\alpha_0}^{\alpha}\frac{d\eta}{l(\eta)}-\frac{K_1}{M_1^2 l_\alpha^2}\int_{\Omega_{\alpha_0}}\rho q^2\,dx
\end{aligned} \tag{2.18}
$$

where l_α denotes $\inf_{\eta\in[\alpha_0,\alpha]} l(\eta)$. We are thus led, for bounded Ω, to the inequality

$$
\int_{\Omega_\alpha}\rho q^2\,dx\leqq\int_{\Omega_{\alpha_0}}\rho q^2\,dx\exp\left\{\frac{2m_1K_1\pi}{M_1^2l_\alpha^2}\int_{\Omega_{\alpha_0}}\rho q^2\,dx\right\}\exp\left\{\frac{-2m_1\pi}{M_1\gamma}\int_{\alpha_0}^{\alpha}\frac{d\eta}{l(\eta)}\right\}, \tag{2.19}
$$

which is the desired decay result. Thus we have established Theorem 1.

THEOREM 1. *If Ω is bounded, then the decay estimate* (2.19) *holds for $\alpha_0<\alpha<\alpha_1$.*

To make the inequality explicit we of course need a bound for $\int_{\Omega_{\alpha_0}}\rho q^2\,dx$. Techniques for finding such bounds have been indicated in [3].

We now consider the case in which Ω is unbounded. From (2.14) we again conclude that if $F(\alpha)$ is positive for some $\tilde{\alpha}$ then it must be positive for all $\alpha > \tilde{\alpha}$. Let us suppose now that $F(\tilde{\alpha}) > 0$ and that

$$l(\eta) \leqq L < \infty. \tag{2.20}$$

Then if

$$\liminf_{\alpha \to \infty} \frac{\int_{S_\alpha} (\rho q^2/|\operatorname{grad} f|)\, ds}{\alpha} = 0 \tag{2.21}$$

we will show that we are led to a contradiction, i.e., it must follow that $F(\alpha) \leqq 0$, for all $\alpha \in (\alpha_0, \infty)$. This can be deduced from the fact that, if $F(\tilde{\alpha}) > 0$, then for $\alpha > \tilde{\alpha}$, (2.7), (2.5), and (2.14) lead to

$$\begin{aligned} \alpha^{-1} \int_{S_\alpha} \frac{\rho q^2}{|\operatorname{grad} f|}\, ds &\geqq \frac{2m_1 \pi F(\tilde{\alpha})}{\alpha\gamma[M_1 l(\alpha) + K_1 \gamma \int_{S_\alpha} (\rho q^2\, ds/|\operatorname{grad} f|)]} \\ &\quad \cdot \exp\left\{ \int_{\tilde{\alpha}}^{\alpha} \frac{2m_1 \pi\, d\eta}{\gamma[M_1 l(\eta) + K_1 \gamma \int_{S_\eta} (\rho q^2\, ds/|\operatorname{grad} f|)]} \right\}. \end{aligned} \tag{2.22}$$

But given any $\varepsilon > 0$ it follows from assumption (2.21) that for sufficiently large α

$$\int_{S_\alpha} \frac{\rho q^2}{|\operatorname{grad} f|}\, ds \leqq \varepsilon\alpha. \tag{2.23}$$

Thus for sufficiently large α

$$\begin{aligned} \varepsilon \geqq \alpha^{-1} \int_{S_\alpha} \frac{\rho q^2}{|\operatorname{grad} f|}\, ds &\geqq \frac{2m_1 \pi F(\tilde{\alpha})}{\alpha\gamma[M_1 L + \gamma K_1 \varepsilon\alpha]} \left[\frac{M_1 L + \gamma K_1 \varepsilon\alpha}{M_1 L + \gamma K_1 \varepsilon\tilde{\alpha}} \right]^{2m_1 \pi / K_1 \gamma^2 \varepsilon} \\ &\geqq 2m_1 K_1 \varepsilon \pi F(\tilde{\alpha}) [M_1 L + \gamma K_1 \varepsilon\alpha]^{[(2m_1 \pi / K_1 \gamma^2 \varepsilon) - 2]} [M_1 L + \gamma K_1 \varepsilon\tilde{\alpha}]^{-2m_1 \pi / K_1 \gamma^2 \varepsilon}. \end{aligned} \tag{2.24}$$

Letting $\alpha \to \infty$ we observe that for ε sufficiently small the right-hand side tends to infinity. Thus we have established Theorem 2.

THEOREM 2. *If Ω is unbounded and $l(\alpha)$ is bounded for all α, then either* $\liminf_{\alpha \to \infty} [\alpha^{-1} \int_{S_\alpha} (\rho q^2/|\operatorname{grad} f|)\, ds] > K > 0$ *or* $F(\alpha) \leqq 0$ *for all* $\alpha \in (\alpha_0, \infty)$.

If $l(\alpha) \to \infty$ as $\alpha \to \infty$, then similar alternative theorems can be derived depending on the order of $l(\alpha)$ as $\alpha \to \infty$. For instance, if the family of curves $x_1 = \alpha$ is used and $l(\alpha) \leqq C\alpha$ for some positive constant C then $\gamma \equiv 1$, and in (2.22) the exponential term is bounded by

$$\begin{aligned} &\exp\left\{ 2m_1 \pi \int_{\tilde{\alpha}}^{\alpha} \frac{d\eta}{[M_1 l(\eta) + K_1 \int_{S_\eta} (\rho q^2\, ds/|\operatorname{grad} f|)]} \right\} \\ &\qquad \leqq \left[\frac{\alpha}{\tilde{\alpha}} \right]^{2m_1 \pi / [M_1 C + K_1 \varepsilon]}. \end{aligned} \tag{2.25}$$

Thus (2.22), (2.23) imply in this case

$$\begin{aligned} \varepsilon &\geqq \alpha^{-1} \int_{S_\alpha} \rho q^2\, ds \\ &\geqq \frac{2m_1 \pi F(\tilde{\alpha})}{[M_1 C + K_1 \varepsilon]} [\tilde{\alpha}]^{-2m_1 \pi / (M_1 C + K_1 \varepsilon)} [\alpha]^{2[m_1 \pi / (M_1 C + K_1 \varepsilon) - 1]}. \end{aligned} \tag{2.26}$$

We observe that if

$$m_1 \pi / M_1 C > 1, \tag{2.27}$$

then we may choose ε so small that the exponent of α is positive. Then, letting $\alpha \to \infty$, we would again be led to a contradiction. We see, therefore, that if $l(\alpha) \leqq C\alpha$ then either $\liminf_{\alpha\to\infty}[\alpha^{-1}\int_{S_\alpha} \rho q^2\, ds] > K > 0$ or $F(\alpha) \leqq 0$ for all $\alpha \in (\alpha_0, \infty)$ provided (2.27) is satisfied.

Alternatively we could have chosen the family $|\mathbf{x}| = \alpha$, in which case the assumption $l(\alpha) \leqq C\alpha$ would have led to the same result. We could also have shown by similar arguments that if $m_1\pi/M_1C > 2$ then either $\liminf_{\alpha\to\infty}\{\alpha^{-2}\int_{S_\alpha} \rho q^2\, ds\} > K > 0$ or $F(\alpha) \leqq 0$ for all $\alpha \in (\alpha_0, \infty)$.

We assume henceforth that conditions as $\alpha \to \infty$ imply that $F(\alpha) \leqq 0$ for all $\alpha > \alpha_0$. It follows then from (2.16) that

$$\int_{\Omega_\alpha} \rho q^2\, dx \leqq \int_{\Omega_{\alpha_0}} \rho q^2\, dx \exp\left\{-2m_1\pi \int_{\alpha_0}^{\alpha} [B(\eta)]^{-1}\, d\eta\right\}, \tag{2.28}$$

and we conclude as before that the decay estimate (2.19) holds now for unbounded Ω.

Suppose that Ω is a wedge of angle 2δ. Then by using $f(x_1, x_2) = |\mathbf{x}|$ in (2.19), we find that, as $\alpha \to \infty$, the energy term $\int_{\Omega_\alpha} \rho q^2\, dx$ tends to zero like $(x_1^2 + x_2^2)^{-\beta}$. The constant β is given by

$$\beta = m_1\pi/2M_1\delta. \tag{2.29}$$

(Note that when $f = |\mathbf{x}|$, γ may be taken to be 1.) It follows that if $m_1 = M_1$, the estimated decay rate β will be the same as the optimal decay rate for harmonic functions (see, e.g., [2], [5], [6]). Since (1.3) yields $m_1 \leqq M_1$, the estimated decay rate (2.29) is always less than or equal to the actual decay rate for harmonic functions.

For the wedge we have $l(\alpha) = 2\delta\alpha$, and if $m_1 = M_1$, condition (2.27) implies that the alternative holds for $\delta < \pi/2$. For angles $\pi/2 < \delta < \pi$ we must replace the condition

$$\liminf_{\alpha\to\infty}\left[\alpha^{-1}\int_{S_\alpha} \frac{\rho q^2\, ds}{|\operatorname{grad} f|}\right] = 0 \quad \text{by} \quad \liminf_{\alpha\to\infty}\left[\alpha^{-p}\int_{S_\alpha} \frac{\rho q^2\, ds}{|\operatorname{grad} f|}\right] = 0$$

for an appropriately chosen $p \in (0, 1)$.

When Ω is a semi-infinite strip, we may take $f(x_1, x_2) = x_1$ and recover from (2.19) the *exponential* decay estimate (2.1) of [3].

3. Decay estimates in the second case. Again in this case we define $F(\alpha)$ as in (2.2) and seek to determine a $k(\alpha)$ such that (2.6) is satisfied. To bound the second integral on the right-hand side of (2.9) we now note that, on using the left-hand inequality in (1.4), we have

$$\int_{S_\alpha} \rho u^2 |\operatorname{grad} f|\, ds \leqq \frac{1}{m_2^2}\int_{S_\alpha} \rho^{-1} u^2 |\operatorname{grad} f|\, ds. \tag{3.1}$$

Thus, on setting

$$\tau = \int_0^s \rho^{-1}|\operatorname{grad} f|\, ds, \qquad \mathscr{B}(\alpha) = \int_{S_\alpha} \rho^{-1}|\operatorname{grad} f|\, ds \tag{3.2}$$

we are now led to

$$\int_{S_\alpha} \rho u^2 |\operatorname{grad} f|\, ds \leqq \frac{\mathscr{B}^2(\alpha)}{\pi^2 m_2^2}\int_{S_\alpha} \frac{\rho}{|\operatorname{grad} f|}\left(\frac{\partial u}{\partial s}\right)^2 ds \tag{3.3}$$

which, together with (2.9), (2.5) yields

$$|F(\alpha)| \leqq \frac{\mathscr{B}(\alpha)}{2m_2\pi} F'(\alpha). \tag{3.4}$$

It is clear now that the arguments following (2.13) carry through and that we obtain (2.19) with m_1, M_1, and K_1 replaced by m_2, M_2, and K_2, respectively, provided either that Ω is bounded or that the appropriate behavior at infinity is assumed.

In [3] decay estimates for other norms of the solution have been given. Similar estimates are obtainable for solutions of the problems considered in this paper.

As has been pointed out in the Introduction, for the case of the minimal surface equation we have $m_2 = M_2 = K_2 = 1$, in which case the bound for the decay rate is given according to (2.19) as $\exp\{-2\pi\gamma^{-1}\int_{\alpha_0}^{\alpha} d\eta/l(\eta)\}$, which for a wedge and for a semi-infinite strip agrees with the optimal decay rate for harmonic functions. (We choose $f = |\mathbf{x}|$, $\gamma = 1$ for a wedge; $f = x_1$, $\gamma = 1$ for a semi-infinite strip.)

Acknowledgments. We are grateful to the referees for their comments on an earlier version of this paper.

REFERENCES

[1] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, Second edition, Cambridge University Press, Cambridge, U.K., 1967.

[2] C. O. HORGAN AND J. K. KNOWLES, *Recent developments concerning Saint-Venant's principle*, Adv. in Appl. Mech., 23 (1983), pp. 179-269.

[3] C. O. HORGAN AND L. E. PAYNE, *Decay estimates for second-order quasilinear partial differential equations*, Adv. in Appl. Math., 5 (1984), pp. 309-332.

[4] ———, *Decay estimates for a class of second-order quasilinear equations in three dimensions*, Arch. Rational Mech. Anal., 86 (1984), pp. 279-289.

[5] O. A. OLEINIK AND G. A. YOSIFIAN, *Boundary value problems for second-order elliptic equations in unbounded domains and Saint-Venant's principle*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 4 (1977), pp. 269-290.

[6] C.-H. WU, *A Saint-Venant principle for the Neumann problem with a non-thin two dimensional domain*, Internat. J. Engrg. Sci., 8 (1970), pp. 389-402.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 789-815, July 1989

003

DECAY ESTIMATES IN STEADY PIPE FLOW*

K. A. AMES† AND L. E. PAYNE‡

Abstract. In this paper St. Venant type results are derived for the flow of viscous fluid in a pipe of arbitrary cross section. In the spirit of earlier work of Horgan and Wheeler [*SIAM J. Appl. Math.*, 35 (1978), pp. 97-116], the decay to fully developed flow as a function of distance from the entry section is investigated. Here, it is not assumed that the flow is fully developed at the exit section. Weighted energy inequalities are derived that lead to estimates for the "energy" associated with the velocity field represented by the difference between the entrance flow and the fully developed flow in a portion of the pipe near the exit section. The analysis is based on a variety of differential inequality techniques and Payne's investigation of uniqueness criteria for steady-state solutions of the Navier-Stokes equations [*Simpos. Internaz. Appl. Fis. Mat.*, 1965, pp. 130-153].

Key words. pipe flow, Navier-Stokes equations, explicit decay estimates

AMS(MOS) subject classifications. 35Q10, 35B45, 76D05

1. Introduction.

Consider the problem of steady flow of a viscous fluid in a pipe of arbitrary smooth cross section. The flow is governed by the Navier-Stokes equations together with the assumption of adherence at the pipe boundary. If the Reynolds number is sufficiently small we expect that, irrespective of the entrance velocity profile, the flow will approach the fully developed flow near the pipe exit if the pipe is sufficiently long. In a recent paper of Horgan and Wheeler [7] the goal of the authors was to show that under certain assumptions this decay to fully developed flow (in the appropriate measure) was exponential. Their results were based on differential inequality techniques developed by Knowles [8] and Toupin [17] in their investigations of St. Venant's Principle in classical elasticity theory (see also Horgan and Knowles [5]).

In the present paper we again address this flow problem with an analysis that has many features in common with that used by Horgan and Wheeler and thus relies heavily on differential inequality techniques. However, our formulation of the problem and methodology differ somewhat from theirs with the consequence that we have obtained slightly better, more explicit estimates for a weighted energy integral associated with the flow.

One of the assumptions that Horgan and Wheeler made in their work was that the entrance flow had already evolved completely into the fully developed flow at the exit end of the pipe and thus that the tangential components of the velocity in the outlet cross section were both zero. This is a somewhat unrealistic assumption since we cannot expect complete evolution in a pipe of finite length. In this paper we relax that assumption, supposing instead that the tangential velocities and their derivatives in the exit section are small. Our assumption has the effect of making the problem considerably more complicated, and we find it convenient to compare the entrance flow to the fully developed flow indirectly through the introduction of a linearized Stokes flow that enters the pipe with a velocity field equivalent to the fully developed one. Such an approach has necessitated the introduction of some additional restrictions on the boundary data for us to obtain our estimates.

* Received by the editors August 1, 1988; accepted for publication October 7, 1988. This research was supported in part by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University and in part by National Science Foundation grant DMS-8600250.

† Department of Mathematical Sciences, Rice University, P.O. Box 1892, Houston, Texas 77251. This work was done while this author was a visiting scientist at the Mathematical Sciences Institute, Cornell University, Ithaca, New York 14853.

‡ Department of Mathematics, Cornell University, Ithaca, New York 14853.

Our main result is an explicit inequality that gives an indication of the spatial development of the entrance flow as it moves through the pipe. If the data and ν are suitably restricted, the inequality we derive establishes a decay estimate for a weighted functional defined on the difference between the entrance velocity field and the Stokes flow. Combining this result with estimates on this functional defined on the difference between the Stokes and fully developed flows, we obtain the desired inequality comparing the entrance and fully developed velocity fields. In fact we show that a certain weighted energy functional of this difference is bounded by the sum of two terms—one a decaying exponential and the other a function of difference in exit data.

In the course of our analysis, it becomes necessary for us to find bounds for the energy of the corresponding linearized Stokes problem. The arguments we used to accomplish this are patterned after those used by Payne [11] in his investigation of uniqueness criteria for steady state solutions of the Navier-Stokes equations. As we will see, our estimates are explicit in the sense that they depend only on ν, the boundary data, and the geometry of the domain.

It should be pointed out that questions related to those discussed in this paper have also been considered by Amick [1], [2], Yosifian [19], [20], and Oleinik [10]. In these latter papers the authors have considered domains with boundaries extending to infinity, concentrating primarily on questions of existence, uniqueness, and regularity. Yosifian [19] has established an energy decay result for the Stokes system (which is of precisely the same form as the equations of incompressible linear elasticity), and Amick [2] has established exponential energy decay rates, but for general cross sections these were not explicitly expressed in terms of the data and geometry.

In the next section, we will formulate the boundary value problem that serves as the basis for our analysis. Section 3 is devoted to a summary of the inequalities and auxiliary results that we will utilize to generate our estimates. We then compare the entrance flow to a Stokes flow whose boundary data on the inlet cross section coincide with that of the fully developed flow. Our main differential inequality is established in § 4 and bounds for the Stokes flow are derived in §§ 5 and 6. Comparison of the entrance and fully developed flows and a discussion of our results, in particular the criteria that ensure decay, are the subjects of § 7. In this section, we also obtain an upper bound for the maximum speed of the fully developed flow in terms of the prescribed net inflow and the domain geometry by taking advantage of results for the analogous Saint-Venant torsion problem. The section also contains some remarks about the constants appearing in our estimates. Finally, part of § 7 is devoted to a brief description of how bounds for the total weighted energy in terms of data and geometry can be derived. We note that because of the tedious nature of the analysis used in this problem, we have relegated some of the details to three appendices.

2. Formulation of problem. In this section we formulate the boundary value problem that provides the framework for our investigation of the flow development of an incompressible viscous fluid in a pipe. Much of our notation coincides with that used by Horgan and Wheeler [7]. We let R denote the interior of a three-dimensional cylindrical pipe of length l and ∂R its boundary. A plane cross section of the pipe with Cartesian coordinates (x_1, x_2, z), $z \in [0, l]$ fixed, will be denoted by S_z and its boundary by ∂S_z. In particular, S_0 represents the inlet cross section and S_l the outlet cross section of the pipe.

The velocity field $u_i(x_1, x_2, x_3)$ $(i = 1, 2, 3)$ and the pressure $p(x_1, x_2, x_3)$ of the fluid are assumed to be classical solutions of the following boundary value problem:

$$\nu \Delta u_i = p_{,i} + u_j u_{i,j} \quad \text{in } R, \tag{2.1}$$

$$u_{i,i} = 0 \quad \text{in } R, \tag{2.2}$$

$$u_i = 0 \quad \text{on } \partial R \backslash (S_0 \cup S_l), \tag{2.3}$$

$$u_i = f_i(x_1, x_2) \quad \text{on } S_0, \tag{2.4}$$

$$u_i = g_i(x_1, x_2) \quad \text{on } S_l. \tag{2.5}$$

Here Δ denotes the Laplace operator and ν is the constant kinematic viscosity. The prescribed entrance profile $f_i(x_1, x_2)$ and exit velocity $g_i(x_1, x_2)$ are assumed to be zero on ∂S_0 and ∂S_l, respectively. The vector field $v_i(x_1, x_2, x_3)$ is the solution of the Stokes problem

$$\nu \Delta v_i = q_{,i} \quad \text{in } R, \tag{2.6}$$

$$v_{i,i} = 0 \quad \text{in } R, \tag{2.7}$$

$$v_i = 0 \quad \text{on } \partial R \backslash (S_0 \cup S_l), \tag{2.8}$$

$$v_i = V\delta_{3i} \quad \text{on } S_0, \tag{2.9}$$

$$v_i = u_i \quad \text{on } S_l. \tag{2.10}$$

Here $(0, 0, V(x_1, x_2))$ represents the fully developed velocity field corresponding to the net inflow

$$\int_{S_0} f_3 \, dx = Q \tag{2.11}$$

and can thus be characterized as the solution of the boundary value problem

$$\nu V_{,\alpha\alpha} = \tilde{p}_{,3} \quad \text{in } S_z, \tag{2.12}$$

$$V = 0 \quad \text{on } \partial S_z, \tag{2.13}$$

$$\int_{S_z} V \, dx = Q. \tag{2.14}$$

The gradient of the pressure $\tilde{p}$ in (2.12) has the form $\tilde{p}_{,i} = -P\delta_{3i}$ where P is a positive constant.

If the velocity vector $g_i(x_1, x_2)$ at the exit end of the pipe is close to that of fully developed flow then the data $g_i - V\delta_{3i}$ may be expected to be small as elements of $W_2^1(S_l)$. We shall have more to say about this later.

In the previous equations as well as the subsequent analysis, we adopt the summation convention and denote differentiation by a comma. Latin subscripts will range from 1-3 while Greek subscripts will range only from 1-2.

Recall that our goal is to compare the entrance flow u_i to the fully developed flow V. To do so we introduce a Stokes flow v_i that coincides with V on the inlet cross section of the pipe and with u_i on the outlet section. We shall first compare u_i and v_i and then v_i and $V\delta_{3i}$. These separate comparisons will lead to estimates for the energy of the difference between the entrance and fully developed flows. As mentioned in the Introduction, our reason for taking this approach rather than comparing u_i and $V\delta_{3i}$ directly as Horgan and Wheeler did, is motivated by the fact that we do not expect the entrance velocity field to evolve completely into the fully developed field in a pipe of finite length. The present approach allows us to avoid making the assumption that $u_\alpha = 0$ and $u_3 = V$ on S_l.

To relate the solution u_i of (2.1)-(2.5) to the solution v_i of (2.6)-(2.10), we define $w_i = u_i - v_i$ and $s = p - q$. Then the boundary value problem governing the difference fields is

$$\nu \Delta w_i = s_{,i} + (w_j + v_j)(w_{i,j} + v_{i,j}) \quad \text{in } R, \tag{2.15}$$

$$w_{i,i} = 0 \quad \text{in } R, \tag{2.16}$$

$$w_i = 0 \quad \text{on } \partial R \backslash S_0, \tag{2.17}$$

$$w_i = f_i - V\delta_{3i} \quad \text{on } S_0. \tag{2.18}$$

In view of (2.11) and (2.14), we have $\int_{S_0} w_3 \, dx = 0$. From this fact as well as equations (2.16)-(2.18) and the divergence theorem, we readily obtain the condition of zero net axial flow, i.e.,

$$\int_{S_z} w_3 \, dx = 0 \quad \text{for } 0 \leqq z \leqq l. \tag{2.19}$$

As pointed out in [7], an analogy between this condition and one that arises in Saint-Venant's Principle of elasticity theory can be made.

3. Auxiliary results. To establish the main estimate of this paper, we have relied on a number of standard inequalities that we summarize in this section.

Let S be a plane domain with boundary ∂S. If $w = 0$ on ∂S, then we have the Poincaré inequality

$$\lambda_1 \int_S w^2 \, dx \leqq \int_S w_{,\alpha} w_{,\alpha} \, dx \tag{3.1}$$

where λ_1 is the smallest positive eigenvalue of

$$\Delta \varphi + \lambda \varphi = 0 \quad \text{in } S, \qquad \varphi = 0 \quad \text{on } \partial S.$$

If, however, $w = 0$ on ∂S and $\int_S w \, dx = 0$, then

$$\lambda_2 \int_S w^2 \, dx \leqq \int_S w_{,\alpha} w_{,\alpha} \, dx \tag{3.2}$$

where λ_2 is the smallest positive eigenvalue of the problem

$$\Delta \varphi + \lambda \varphi = \hat{k} \quad \text{in } S, \qquad \varphi = 0 \quad \text{on } \partial S,$$

$$\int_S \varphi \, dx = 0$$

for a constant $\hat{k}$. It is easily seen that $\lambda_1 \leqq \lambda_2$. In fact, as indicated in § 5, a sharper lower bound for λ_2 can be found.

In addition to inequalities (3.1) and (3.2), we will make use of the following two Sobolev inequalities that hold for any sufficiently smooth function w with compact support in either $\mathbb{R}^2$ or $\mathbb{R}^3$:

$$\text{(i)} \quad \int\int_{-\infty}^{\infty} w^4 \, dx \leqq \frac{1}{2} \left(\int\int_{-\infty}^{\infty} w^2 \, dx \right) \left(\int\int_{-\infty}^{\infty} w_{,\alpha} w_{,\alpha} \, dx \right), \tag{3.3}$$

$$\text{(ii)} \quad \int\int\int_{-\infty}^{\infty} w^6 \, dx \leqq \Omega \left(\int\int\int_{-\infty}^{\infty} w_{,j} w_{,j} \, dx \right)^3. \tag{3.4}$$

In a number of papers [9], [16], the best constant in (3.4) has been computed to have the value $\Omega = (1/27)(2/\pi)^4$.

An important result that will be needed in deriving our estimates is contained in the following theorem that is given in Babuška and Aziz [3] under less stringent smoothness hypotheses.

THEOREM A. *Let S be a plane Lipschitz domain and let w be a function that is continuously differentiable on $\bar{S}$ and satisfies $\int_S w\,dx = 0$. Then there exists a vector function ψ_α such that*

$$\psi_{\alpha,\alpha} = w \quad \text{in } S, \qquad \psi_\alpha = 0 \quad \text{on } \partial S \tag{3.5}$$

and a positive constant C depending only on the geometry of S such that

$$\int_S \psi_{\alpha,\beta}\psi_{\alpha,\beta}\,dx \leqq C \int_S (\psi_{\alpha,\alpha})^2\,dx. \tag{3.6}$$

This theorem asserts the existence of such a function ψ_α, which clearly is not unique. In fact, in our applications we will not be interested in the function ψ_α itself (we require only the existence of such a function) but rather in the constant C. More specifically, for a given domain S we would like to find the smallest possible value of C or, barring this, a close upper bound for this optimal C. Since the optimal C does not depend on the size of S but only on its shape (it can be thought of as a dimensionless eigenvalue defined by (3.6)), it can easily be seen that the optimal constant for a circle of arbitrary radius is $C = 1$. In [6] it has been shown how this optimal constant is related to an optimal Korn's constant and to an optimal Friedrich's constant. In this latter paper an explicit upper bound for the optimal C for a star-shaped region is given. Not only will inequality (3.6) be useful in establishing the final estimate but this result will enable us to eliminate the pressure difference s via the introduction of an auxiliary function ψ_α, the existence of which is ensured by the theorem since w_3 satisfies its hypotheses.

4. Comparison of u_i and v_i. We will divide our analysis into two parts. In the first part, we derive a first-order inequality for a weighted energy integral that is defined for solutions of (2.15)-(2.18). Integration of this inequality results in the desired estimates. In the course of this derivation, we shall need some inequalities for the energy associated with the linearized Stokes problem that corresponds to (2.15)-(2.18). Section 5 will be devoted to establishing these inequalities.

We define a weighted energy integral by

$$\Phi(z) = \frac{1}{3}\int_z^l \int_{S_\xi} (\xi - z)^3 w_{i,j} w_{i,j}\,dx\,d\xi \tag{4.1}$$

for solutions w_i of the problem (2.15)-(2.18). In this section, we show that $\Phi(z)$ satisfies a first-order differential inequality of the form

$$K\frac{d\Phi}{dz} + M\Phi \leqq M_2 l^3 Q_0 \tag{4.2}$$

where K, M, and M_2 are positive constants and Q_0 is a data term. Integration of this inequality with the appropriate data assumptions leads to a result of the form

$$\Phi(z) \leqq \Phi(0)e^{-\kappa_1 z} + \kappa_2 l^3 Q_0(1 - e^{-\kappa_1 z}) \tag{4.3}$$

for $0 \leqq z \leqq l$. Here $\kappa_1 = M/K$ and $\kappa_2 = M_2/M$.

To obtain (4.2), we first observe that the function $\Phi(z)$ can be rewritten after an integration by parts and substitution of the differential equations for w_i as

$$\begin{aligned}\Phi(z) = &-\int_z^l \int_{S_\xi} (\xi - z)^2 w_i w_{i,3}\, dx\, d\xi + \frac{1}{\nu}\int_z^l \int_{S_\xi} (\xi - z)^2 w_3 s\, dx\, d\xi \\ &+\frac{1}{2\nu}\int_z^l \int_{S_\xi} (\xi - z)^2 w_i w_i v_3\, dx\, d\xi \\ &-\frac{1}{3\nu}\int_z^l \int_{S_\xi} (\xi - z)^3 w_i [w_j w_{i,j} + (w_j + v_j) v_{i,j}]\, dx\, d\xi.\end{aligned} \tag{4.4}$$

From the definition of $\Phi(z)$ in (4.1), we also have

$$\frac{d\Phi}{dz} = -\int_z^l \int_{S_\xi} (\xi - z)^2 w_{i,j} w_{i,j}\, dx\, d\xi. \tag{4.5}$$

To derive the desired differential inequality, we need to eliminate the pressure term s from the expression (4.4) for $\Phi(z)$. This can be accomplished with the aid of Theorem A from § 3. We thus introduce the vector function ψ_α that satisfies

$$\psi_{\alpha,\alpha} = w_3 \quad \text{in } S_z, \tag{4.6}$$

$$\psi_\alpha = 0 \quad \text{on } \partial S_z \tag{4.7}$$

for $z \in [0, l]$. The existence of such a function is guaranteed by Theorem A since $\int_{S_z} w_3\, dx = 0$ for $0 \leqq z \leqq l$. If we now multiply the first two of equations (2.14) by $(\xi - z)^2 \psi_\alpha$ and integrate over the domain, we obtain (following Horgan and Wheeler [7])

$$\begin{aligned}&\int_z^l \int_{S_\xi} (\xi - z)^2 s w_3\, dx\, d\xi \\ &\quad = \int_z^l \int_{S_\xi} (\xi - z)^2 [\psi_\alpha (w_j + v_j)(w_{\alpha,j} + v_{\alpha,j}) + \nu \psi_{\alpha,j} w_{\alpha,j} - \nu(\psi_\alpha w_{\alpha,3})_{,3}]\, dx\, d\xi.\end{aligned} \tag{4.8}$$

Substitution of this expression into (4.4) leads to

$$\begin{aligned}\Phi(z) = &-\int_z^l \int_{S_\xi} (\xi - z)^2 w_i w_{i,3}\, dx\, d\xi - \int_z^l \int_{S_\xi} (\xi - z)^2 (\psi_\alpha w_{\alpha,3})_{,3}\, dx\, d\xi \\ &+\int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,j} w_{\alpha,j}\, dx\, d\xi + \frac{1}{\nu}\int_z^l \int_{S_\xi} (\xi - z)^2 \psi_\alpha (w_j + v_j)(w_{\alpha,j} + v_{\alpha,j})\, dx\, d\xi \\ &+\frac{1}{2\nu}\int_z^l \int_{S_\xi} (\xi - z)^2 w_i w_i v_3\, dx\, d\xi - \frac{1}{3\nu}\int_z^l \int_{S_\xi} (\xi - z)^2 w_i (w_j + v_j) v_{i,j}\, dx\, d\xi \\ &-\frac{1}{3\nu}\int_z^l \int_{S_\xi} (\xi - z)^3 w_i w_j w_{i,j}\, dx\, d\xi.\end{aligned} \tag{4.9}$$

We have thus written $\Phi(z)$ as the sum of eleven integrals Φ_k each of which we must now bound in terms of Φ and $d\Phi/dz$. We indicate here how this can be done for four representative Φ_k; the remaining seven integrals are treated in Appendix I. We note that all but Φ_9 and Φ_{10} can be bounded in terms of $d\Phi/dz$ alone. As we shall see, the bounds on Φ_9 and Φ_{10} will generate a decay criterion.

Consider first

$$\Phi_3 = \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,j} w_{\alpha,j}\, dx\, d\xi$$
$$= \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,\beta} w_{\alpha,\beta}\, dx\, d\xi + \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,3} w_{\alpha,3}\, dx\, d\xi.$$

Using Schwarz's inequality and inequalities (3.1), (3.2) and (3.6), we have

$$\Phi_3 \leqq \left(\frac{C}{\lambda_2}\right)^{1/2} \left(\int_z^l \int_{S_\xi} (\xi - z)^2 w_{3,\beta} w_{3,\beta}\, dx\, d\xi \right)^{1/2} \left(\int_z^l \int_{S_\xi} (\xi - z)^2 w_{\alpha,\beta} w_{\alpha,\beta}\, dx\, d\xi \right)^{1/2}$$
$$+ \left(\int_z^l \int_{S_\xi} (\xi - z)^2 w_{\alpha,3} w_{\alpha,3}\, dx\, d\xi \right)^{1/2} \left(\int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,3} \psi_{\alpha,3}\, dx\, d\xi \right)^{1/2}. \tag{4.10}$$

Since $\psi_\alpha = 0$ on ∂S_z, we have $\psi_{\alpha,3} = 0$ on ∂S_z and thus

$$\int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,3} \psi_{\alpha,3}\, dx\, d\xi \leqq \frac{1}{\lambda_1} \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,\beta 3} \psi_{\alpha,\beta 3}\, dx\, d\xi$$
$$\leqq \frac{C}{\lambda_1} \int_z^l \int_{S_\xi} (\xi - z)^2 (\psi_{\alpha,\alpha 3})^2\, dx\, d\xi$$
$$= \frac{C}{\lambda_1} \int_z^l \int_{S_\xi} (\xi - z)^2 w_{3,3}^2\, dx\, d\xi$$

after application of inequalities (3.1) and (3.6). From the arithmetic-geometric mean inequality, we then obtain the bound

$$\Phi_3 \leqq -\frac{1}{2} \sqrt{\frac{C}{\lambda_1}} \frac{d\Phi}{dz}. \tag{4.11}$$

To bound $\Phi_5 = 1/\nu \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_\alpha v_j w_{\alpha,j}\, dx\, d\xi$, we first rewrite the integral as

$$\Phi_5 = \frac{1}{\nu} \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_\alpha w_{\alpha,3} V\, dx\, d\xi + \frac{1}{\nu} \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_\alpha w_{\alpha,j} \zeta_j\, dx\, d\xi = I_1 + I_2$$

where $\zeta_j = v_j - V\delta_{3j}$. The first integral is bounded by applying the Schwarz and Poincaré inequalities. We have

$$I_1 \leqq \frac{v_s}{\nu \sqrt{\lambda_1}} \left(\int_z^l \int_{S_\xi} (\xi - z)^2 w_{\alpha,3} w_{\alpha,3}\, dx\, d\xi \right)^{1/2} \left(\int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,\beta} \psi_{\alpha,\beta}\, dx\, d\xi \right)^{1/2}$$

where $v_s = \max_{S_z} |V(x_1, x_2)|$. It then follows from (3.6), (3.2), and the arithmetic-geometric mean inequality that

$$I_1 \leqq -\frac{v_s}{2\nu} \sqrt{\frac{C}{\lambda_1 \lambda_2}} \frac{d\Phi}{dz}. \tag{4.12}$$

A bound for I_2 is established through the use of Schwarz's inequality and the inequalities (3.1)-(3.3) as well as (3.6). We obtain

$$I_2 \leqq \frac{1}{2\nu} \sqrt{\frac{C}{\lambda_2}} \left(\frac{1}{2\lambda_1}\right)^{1/4} \max_z \left(\int_{S_z} (\zeta_i \zeta_i)^2\, dx \right)^{1/4}$$
$$\cdot \left[\int_z^l \int_{S_\xi} (\xi - z)^2 w_{3,\beta} w_{3,\beta}\, dx\, d\xi + \int_z^l \int_{S_\xi} (\xi - z)^2 w_{\alpha,j} w_{\alpha,j}\, dx\, d\xi \right].$$

Consequently, we conclude that

$$\Phi_5 \leqq -\frac{v_s}{2\nu}\sqrt{\frac{C}{\lambda_1\lambda_2}}\frac{d\Phi}{dz} - \frac{1}{2\nu}\sqrt{\frac{C}{\lambda_2}}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}\frac{d\Phi}{dz}. \tag{4.13}$$

Consider now

$$\Phi_9 = -\frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3 w_3 w_\alpha V_{,\alpha}\,dx\,d\xi - \frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3 w_i w_j \zeta_{i,j}\,dx\,d\xi = J_1 + J_2$$

with ζ_i defined as before. Integration of the first integral leads to

$$J_1 = \frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3 w_\alpha w_{3,\alpha} V\,dx\,d\xi + \frac{1}{2\nu}\int_z^l\int_{S_\xi}(\xi-z)^2 w_3^2 V\,dx\,d\xi.$$

From Schwarz's inequality, (3.1), (3.2), and the arithmetic-geometric mean inequality, we obtain the bound

$$J_1 \leqq \frac{1}{3}\left(\frac{v_s}{2\nu\sqrt{\lambda_1}}\right)\left[\int_z^l\int_{S_\xi}(\xi-z)^3 w_{3,\alpha}w_{3,\alpha}\,dx\,d\xi + \int_z^l\int_{S_\xi}(\xi-z)^3 w_{\alpha,\beta}w_{\alpha,\beta}\,dx\,d\xi\right]$$
$$+\frac{v_s}{2\nu\lambda_2}\int_z^l\int_{S_\xi}(\xi-z)^2 w_{3,\beta}w_{3,\beta}\,dx\,d\xi.$$

Thus,

$$J_1 \leqq \frac{v_s}{2\nu\sqrt{\lambda_1}}\Phi - \frac{v_s}{2\nu\lambda_2}\frac{d\Phi}{dz}. \tag{4.14}$$

Turning to J_2, we observe that it may be rewritten as

$$J_2 = \frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3 w_{i,j}w_j\zeta_i\,dx\,d\xi + \frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi-z)^2 w_i w_3\zeta_i\,dx\,d\xi.$$

It then follows by applying Schwarz's inequality, (3.1)-(3.3), and the arithmetic-geometric mean inequality in the proper sequence that

$$J_2 \leqq \frac{1}{3\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}\int_z^l\int_{S_\xi}(\xi-z)^3 w_{i,j}w_{i,j}\,dx\,d\xi$$
$$+\frac{1}{2\nu\sqrt{\lambda_2}}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}$$
$$\cdot\left[\int_z^l\int_{S_\xi}(\xi-z)^2 w_{i,\beta}w_{i,\beta}\,dx\,d\xi + \int_z^l\int_{S_\xi}(\xi-z)^2 w_{3,\beta}w_{3,\beta}\,dx\,d\xi\right].$$

Hence, we see that

$$J_2 \leqq \frac{1}{\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}\Phi - \frac{1}{\nu\sqrt{\lambda_2}}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}\frac{d\Phi}{dz}. \tag{4.15}$$

Combining (4.14) and (4.15), we have a bound for Φ_9 in terms of Φ and $d\Phi/dz$.

Finally let us consider Φ_{10}, which we rewrite as the sum of three integrals:

$$\Phi_{10} = -\frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3 w_i\zeta_i\zeta_{i,j}\,dx\,d\xi - \frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3 w_i\zeta_{i,3}V\,dx\,d\xi$$
$$-\frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3 w_3\zeta_\alpha V_{,\alpha}\,dx\,d\xi$$
$$= K_1 + K_2 + K_3.$$

We see that

$$K_1+K_2\leqq\frac{1}{3\nu}\int_z^l(\xi-z)^3\left(\int_{S_\xi}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}\left(\int_{S_\xi}(w_iw_i)^2\,dx\right)^{1/4}\left(\int_{S_\xi}\zeta_{i,j}\zeta_{i,j}\,dx\right)^{1/2}d\xi$$
$$+\frac{v_s}{3\nu}\int_z^l(\xi-z)^3\left(\int_{S_\xi}w_iw_i\,dx\right)^{1/2}\left(\int_{S_\xi}\zeta_{i,3}\zeta_{i,3}\,dx\right)^{1/2}d\xi$$

from which, after applying the appropriate inequalities, it follows that

$$K_1+K_2\leqq\frac{1}{3\nu}\frac{1}{2}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}$$
$$\cdot\left[a_1\int_z^l\int_{S_\xi}(\xi-z)^3w_{i,\beta}w_{i,\beta}\,dx\,d\xi+\frac{1}{a_1}\int_z^l\int_{S_\xi}(\xi-z)^3\zeta_{i,j}\zeta_{i,j}\,dx\,d\xi\right] \tag{4.16}$$
$$+\frac{1}{3}\frac{v_s}{2\nu\sqrt{\lambda_1}}\left[a_2\int_z^l\int_{S_\xi}(\xi-z)^3w_{i,\beta}w_{i,\beta}\,dx\,d\xi+\frac{1}{a_2}\int_z^l\int_{S_\xi}(\xi-z)^3\zeta_{i,3}\zeta_{i,3}\,dx\,d\xi\right]$$

where a_1 and a_2 are arbitrary positive constants. We now observe that integration, the fact that $\int_{S_z}\zeta_3\,dx=0$, and the divergence free condition allow us to express K_3 as

$$K_3=\frac{1}{3\nu}\int_z^l\int_{S_\xi}(\xi-z)^3w_{3,j}\zeta_jV\,dx\,d\xi+\frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi-z)^3w_3\zeta_3V\,dx\,d\xi$$
$$\leqq\frac{1}{3}\frac{v_s}{2\nu\sqrt{\lambda_1}}\left[a_2\int_z^l\int_{S_\xi}(\xi-z)^3w_{3,j}w_{3,j}\,dx\,d\xi+\frac{1}{a_2}\int_z^l\int_{S_\xi}(\xi-z)^3\zeta_{j,\beta}\zeta_{j,\beta}\,dx\,d\xi\right] \tag{4.17}$$
$$+\frac{v_s}{2\nu\lambda_2}\left[\int_z^l\int_{S_\xi}(\xi-z)^2w_{3,\beta}w_{3,\beta}\,dx\,d\xi+\int_z^l\int_{S_\xi}(\xi-z)^2\zeta_{3,\beta}\zeta_{3,\beta}\,dx\,d\xi\right].$$

From (4.16) and (4.17), we thus conclude that

$$\Phi_{10}\leqq\left\{\frac{a_1}{2\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}+\frac{a_2v_s}{\nu\sqrt{\lambda_1}}\right\}\Phi$$
$$+\frac{v_s}{2\nu\lambda_2}\left[-\frac{d\Phi}{dz}+\int_z^l\int_{S_\xi}(\xi-z)^2\zeta_{i,j}\zeta_{i,j}\,dx\,d\xi\right] \tag{4.18}$$
$$+\left\{\frac{1}{6\nu a_1}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}+\frac{v_s}{3\nu a_2\sqrt{\lambda_1}}\right\}$$
$$\cdot\int_z^l\int_{S_\xi}(\xi-z)^3\zeta_{i,j}\zeta_{i,j}\,dx\,d\xi.$$

In Appendix I, we derive the following inequalities:

$$\Phi_1\leqq\frac{-1}{2\sqrt{\lambda_1}}\frac{d\Phi}{dz}, \tag{4.19}$$

$$\Phi_2\leqq-2\sqrt{\frac{C}{\lambda_1}}\frac{d\Phi}{dz}, \tag{4.20}$$

$$\Phi_4\leqq-\frac{2}{\nu}\sqrt{\frac{C}{\lambda_1}}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(w_\alpha w_\alpha)^2\,dx\right)^{1/4}\frac{d\Phi}{dz}, \tag{4.21}$$

$$\Phi_6\leqq-\frac{\sqrt{C}}{2\nu}\max_z\left(\int_{S_z}(\zeta_\alpha\zeta_\alpha)^2\,dx\right)^{1/4}\left[\frac{1}{\sqrt{\lambda_2}}\left(\frac{1}{2\lambda_1}\right)^{1/4}+\frac{5}{\sqrt{\lambda_1}}\left(\frac{1}{2\lambda_2}\right)^{1/4}\right]\frac{d\Phi}{dz}, \tag{4.22}$$

$$\Phi_7 \leqq \frac{1}{2\nu}\sqrt{\frac{C}{\lambda_2}}\left\{\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}+\frac{v_s}{\sqrt{\lambda_1}}\right\}$$
$$\cdot\left[-\frac{d\Phi}{dz}+\int_z^l\int_{S_\xi}(\xi-z)^2\zeta_{i,j}\zeta_{i,j}\,dx\,d\xi\right], \tag{4.23}$$

$$\Phi_8 \leqq -\frac{1}{2\nu}\left[\frac{1}{2\sqrt{\lambda_1}}\max_z\left(\int_{S_z}\zeta_3^2\,dx\right)^{1/2}+\frac{v_s}{\lambda_1}\right]\frac{d\Phi}{dz}, \tag{4.24}$$

$$\Phi_{11} \leqq -\frac{1}{2\nu\sqrt{2\lambda_1}}\max_z\left(\int_{S_z}w_3^2\,dx\right)^{1/2}\frac{d\Phi}{dz}. \tag{4.25}$$

Combining the relevant inequalities, we obtain the result

$$\Phi \leqq M_1\Phi - K\frac{d\Phi}{dz}+N \tag{4.26}$$

where

$$M_1=\frac{v_s}{2\nu\sqrt{\lambda_1}}(1+2a_2)+\frac{1}{\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}\left(1+\frac{a_1}{2}\right), \tag{4.27}$$

$$N=\left\{\frac{1}{6\nu a_1}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}+\frac{v_s}{3\nu a_2\sqrt{\lambda_1}}\right\}\int_z^l\int_{S_\xi}(\xi-z)^3\zeta_{i,j}\zeta_{i,j}\,dx\,d\xi$$
$$+\left\{\frac{v_s}{2\nu\lambda_2}+\frac{1}{2\nu}\frac{\sqrt{C}}{\sqrt{\lambda_2}}\left[\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}+\frac{v_s}{\sqrt{\lambda_1}}\right]\right\} \tag{4.28}$$
$$\cdot\int_z^l\int_{S_\xi}(\xi-z)^2\zeta_{i,j}\zeta_{i,j}\,dx\,d\xi,$$

and

$$K=\frac{1}{2\sqrt{\lambda_1}}(1+5\sqrt{C})+\frac{v_s}{\nu}\left(\sqrt{\frac{C}{\lambda_1\lambda_2}}+\frac{1}{\lambda_2}+\frac{1}{2\lambda_1}\right)$$
$$+\frac{1}{2\nu\sqrt{2\lambda_1}}\left[\max_z\left(\int_{S_z}w_3^2\,dx\right)^{1/2}+\max_z\left(\int_{S_z}\zeta_3^2\,dx\right)^{1/2}\right]$$
$$+\frac{2}{\nu}\sqrt{\frac{C}{\lambda_2}}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(w_\alpha w_\alpha)^2\,dx\right)^{1/4} \tag{4.29}$$
$$+\frac{\sqrt{C}}{2\nu}\left[\frac{1}{\sqrt{\lambda_2}}\left(\frac{1}{2\lambda_1}\right)^{1/4}+\frac{5}{\sqrt{\lambda_1}}\left(\frac{1}{2\lambda_2}\right)^{1/4}\right]\max_z\left(\int_{S_z}(\zeta_\alpha\zeta_\alpha)^2\,dx\right)^{1/4}$$
$$+\frac{1}{\nu\sqrt{\lambda_2}}\left(\frac{1}{2\lambda_1}\right)^{1/4}\left(1+\frac{3}{2}\sqrt{C}\right)\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}.$$

We thus have the inequality

$$K\frac{d\Phi}{dz}+(1-M_1)\Phi \leqq N.$$

To ensure decay, we require

$$M \equiv 1-M_1 > 0. \tag{4.30}$$

This condition yields a restriction on the flow. We shall subsequently require a somewhat stronger hypothesis if decay is to occur. We shall consider these restrictions more thoroughly in § 7.

5. Bounds in the Stokes flow problem. The definitions of the constants M_1 and K involve the quantities $\max_z (\int_{S_z} w_3^2\,dx)^{1/2}$, $\max_z (\int_{S_z} (w_\alpha w_\alpha)^2\,dx)^{1/4}$, $\max_z (\int_{S_z} \zeta_3^2\,dx)^{1/2}$, and $\max_z (\int_{S_z} (\zeta_i\zeta_i)^2\,dx)^{1/4}$. In addition, we see that N is expressed in terms of the integrals $\int_z^l \int_{S_\xi} (\xi - z)^n \zeta_{i,j}\zeta_{i,j}\,dx\,d\xi$ $(n = 2, 3)$. Since we would like our estimate to depend only on the parameter ν, the boundary data, and the geometry of the domain, we need to obtain bounds for the aforementioned expressions in terms of these quantities. The task of obtaining such estimates ultimately reduces to one of deriving bounds in terms of the geometry and the data for the linearized Stokes problem. This section will be devoted to finding these bounds. We note that our analysis is modeled on that used by Payne [11] in his investigation of uniqueness criteria for steady state solutions of the Navier-Stokes equations. In § 6 we shall address the problem of deriving bounds for the energy associated with the velocity field $\zeta_i = v_i - V\delta_{3i}$.

5.1. Estimates for $\mathscr{L}_2$ and energy integrals. We first observe that since $w_{i,i} = 0$,

$$\int_{S_z} w_3^2\,dx = -2\int_z^l \int_{S_\xi} w_3 w_{3,3}\,dx\,d\xi$$

$$= -(2-a)\int_z^l \int_{S_\xi} w_3 w_{3,3}\,dx\,d\xi - a\int_z^l \int_{S_\xi} w_\alpha w_{3,\alpha}\,dx\,d\xi$$

for any constant a. Application of Schwarz's inequality, (3.1), (3.2), and the arithmetic-geometric mean inequality then lead for $0 < a < 2$ to

$$\int_{S_z} w_3^2\,dx \leqq \frac{ac}{2\sqrt{\lambda_1}}\int_z^l \int_{S_\xi} w_{\alpha,\beta} w_{\alpha,\beta}\,dx\,d\xi + \frac{(2-a)b}{2\sqrt{\lambda_2}}\int_z^l \int_{S_\xi} w_{3,3}^2\,dx\,d\xi$$

$$+\left(\frac{2-a}{2b\sqrt{\lambda_2}} + \frac{a}{2c\sqrt{\lambda_1}}\right)\int_z^l \int_{S_\xi} w_{3,\beta} w_{3,\beta}\,dx\,d\xi \tag{5.1}$$

where b and c are positive constants that we choose so that the coefficients of the three integrals in (5.1) are equal. Then,

$$\int_{S_z} w_3^2\,dx \leqq \frac{\sqrt{\lambda_1(2-a)^2 + \lambda_2 a^2}}{2\sqrt{\lambda_1\lambda_2}}\int_z^l \int_{S_\xi} w_{i,j} w_{i,j}\,dx\,d\xi.$$

The optimal choice $a = 2\lambda_1/(\lambda_1 + \lambda_2)$ leads to the result

$$\int_{S_z} w_3^2\,dx \leqq \frac{1}{\sqrt{\lambda_1} + \sqrt{\lambda_2}}\int_z^l \int_{S_\xi} w_{i,j} w_{i,j}\,dx\,d\xi. \tag{5.2}$$

A bound on $\max_z (\int_{S_z} w_3^2\,dx)^{1/2}$ in terms of $\int_0^l \int_S w_{i,j} w_{i,j}\,dx\,dz$ where S is a generic cross section easily follows from (5.2).

We can also bound $\max_z (\int_{S_z} (w_\alpha w_\alpha)^2\,dx)^{1/4}$ in terms of $\int_0^l \int_S w_{i,j} w_{i,j}\,dx\,dz$ since

$$\int_{S_z} (w_\alpha w_\alpha)^2\,dx = -4\int_z^l \int_{S_\xi} w_\alpha w_{\alpha,3} w_\beta w_\beta\,dx\,d\xi$$

$$\leqq 4\left(\int_z^l \int_{S_\xi} w_{\alpha,3} w_{\alpha,3}\,dx\,d\xi\right)^{1/2}\left(\int_z^l \int_{S_\xi} (w_\alpha w_\alpha)^3\,dx\,d\xi\right)^{1/2}$$

by Schwarz's inequality. Applying the Sobolev inequality (3.4),[1] we obtain

$$\int_{S_z} (w_\alpha w_\alpha)^2 \, dx \leqq 8\Omega^{1/2}\left(\int_0^l \int_{S_\xi} w_{i,j} w_{i,j} \, dx \, dz\right)^2. \tag{5.3}$$

This inequality then leads to the desired bound.

In view of these expressions, we must now obtain a bound for $\int_0^l \int_S w_{i,j} w_{i,j} \, dx \, dz$. To accomplish this, we introduce an auxiliary vector field F_i that satisfies the Stokes problem

$$\begin{aligned} &\nu \Delta F_i = \tilde{q}_{,i} \quad \text{in } R = S \times [0, l], \\ &F_{i,i} = 0 \quad \text{in } R, \\ &F_i = w_i \quad \text{on } \partial R. \end{aligned} \tag{5.4}$$

Our task is now twofold. We shall first compare the energy $\int_0^l \int_S w_{i,j} w_{i,j} \, dx \, dz$ to the energy $\int_0^l \int_S F_{i,j} F_{i,j} \, dx \, dz$ associated with (5.4). Second, we shall obtain a bound on the energy of the Stokes problem in terms of the geometry and boundary data (which are just the data for w_i). In this way, we are led to the desired bound on $\int_0^l \int_S w_{i,j} w_{i,j} \, dx \, dz$ in terms of its data and the geometry of the domain.

Let us begin by noting that since both w_i and F_i are divergence free and $w_i - F_i$ vanishes on ∂R, we have

$$\int_0^l \int_S w_{i,j} w_{i,j} \, dx \, dz - \int_0^l \int_S F_{i,j} F_{i,j} \, dx \, dz = \int_0^l \int_S (w_i - F_i)_{,j} (w_i - F_i)_{,j} \, dx \, dz. \tag{5.5}$$

Applying the divergence theorem to the integral on the right side of (5.5), making use of the governing equations, and adding in the term $(1/\nu) \int_0^l \int_S (w_i - F_i)(w_i - F_i)_{,j} u_j \, dx \, dz$, which clearly gives no contribution, we find that

$$I = \int_0^l \int_S (w_i - F_i)_{,j} (w_i - F_i)_{,j} \, dx \, dz = \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,j} (w_j + v_j)(v_i + F_i) \, dx \, dz.$$

Adding and subtracting the term $(1/\nu) \int_0^l \int_S (w_i - F_i)_{,j} v_i F_j \, dx \, dz$ and introducing $\zeta_i = v_i - V\delta_{3i}$, we obtain I as the sum of nine integrals I_k,

$$\begin{aligned} I = {} & \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,j} (w_j - F_j) \zeta_i \, dx \, dz + \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,j} (w_j - F_j) V\delta_{3i} \, dx \, dz \\ & + \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,j} w_j F_i \, dx \, dz + \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,j} (\zeta_i F_j + \zeta_j F_i) \, dx \, dz \\ & + \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,j} (F_j V\delta_{3i} + F_i V\delta_{3j}) \, dx \, dz \\ & + \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,j} \zeta_i \zeta_j \, dx \, dz + \frac{1}{\nu} \int_0^l \int_S (w_3 - F_3)_{,\alpha} \zeta_\alpha V \, dx \, dz \\ & - \frac{1}{\nu} \int_0^l \int_S (w_3 - F_3) \zeta_{3,3} V \, dx \, dz + \frac{1}{\nu} \int_0^l \int_S (w_i - F_i)_{,3} \zeta_i V \, dx \, dz. \end{aligned} \tag{5.6}$$

Each of the I_k can be bounded in terms of I, $\int_0^l \int_S w_{i,j} w_{i,j} \, dx \, dz$, $\int_0^l \int_S F_{i,j} F_{i,j} \, dx \, dz$ and $\int_0^l \int_S \zeta_{i,j} \zeta_{i,j} \, dx \, dz$. Using these bounds and (5.5), we can obtain a bound for

[1] To apply (3.4) we have extended w_i as an even function across $z = 0$.

$\int_0^l \int_S w_{i,j} w_{i,j}\, dx\, dz$. In Appendix II, we outline how the proper use of inequalities and (5.5) leads us to the inequality

$$\begin{aligned}&\{1-k_1 v_s - k_2 \mathscr{F}^{1/2} - k_3 \mathscr{H}^{1/2}\} \int_0^l \int_S w_{i,j} w_{i,j}\, dx\, dz \\ &\leqq \left\{1 - k_4 v_s - \frac{k_2}{2} \mathscr{F}^{1/2} - k_5 \mathscr{H}^{1/2}\right\} \mathscr{F} + \{k_6 v_s + k_7 \mathscr{H}^{1/2}\} \mathscr{H}\end{aligned} \tag{5.7}$$

where

$$\mathscr{F} = \int_0^l \int_S F_{i,j} F_{i,j}\, dx\, dz \quad \text{and} \quad \mathscr{H} = \int_0^l \int_S \zeta_{i,j} \zeta_{i,j}\, dx\, dz.$$

Here the k_i are positive constants defined by

$$\begin{aligned} k_1 &= \frac{1}{2\nu\sqrt{\lambda_1}} [\sqrt{(\lambda_1+\lambda_2)/\lambda_2}\,(1+\varepsilon_1) + \hat{c}], \\ k_2 &= \frac{1}{\nu}\left(\frac{4}{\lambda_1}\right)^{1/4} \Omega^{1/8}, \qquad k_3 = k_2(1+\delta), \\ k_4 &= k_1 - \frac{2}{\hat{c}\nu\sqrt{\lambda_1}}, \qquad k_5 = k_3 - \frac{2}{\delta} k_2, \\ k_6 &= \frac{1}{\varepsilon_1 \nu\sqrt{\lambda_1}} \sqrt{(\lambda_1+\lambda_2)/\lambda_2}, \qquad k_7 = \frac{1}{2\delta} k_2 \end{aligned} \tag{5.8}$$

where $0 < \varepsilon_1 \ll 1$, $0 < \delta \ll 1$, and $\hat{c}$ is an arbitrary positive constant. Inequality (5.7) is the desired bound on $\int_0^l \int_S w_{i,j} w_{i,j}\, dx\, dz$ provided

$$1 - k_1 v_s - k_2 \mathscr{F}^{1/2} - k_3 \mathscr{H}^{1/2} > 0. \tag{5.9}$$

As we discuss in § 7, this condition generates another restriction on the flow if energy decay is to occur.

5.2. Estimates for the Stokes flow. We now turn to the task of finding a bound for $\int_0^l \int_S F_{i,j} F_{i,j}\, dx\, dz$ in terms of the boundary data and geometry. Here we shall assume in order to make our results explicit that the region $R = S \times [0, l]$ is star-shaped. With a little more effort a bound for the Stokes flow energy can be found for regions that are not star-shaped (see Payne [11]), but we will not pursue this topic in the present paper.

We first observe that

$$\int_R F_{i,j} F_{i,j}\, d\mathscr{V} = \int_R F_{i,j}(F_{i,j} - F_{j,i})\, d\mathscr{V} - \int_{S_0} (F_{3,\alpha} F_\alpha - F_{\alpha,\alpha} F_3)\, dx, \tag{5.10}$$

since F_i is divergence free and vanishes everywhere on ∂R except on S_0. Consequently, to obtain the desired bound we need to estimate the first integral on the right side of (5.10). Let $y_k = x_k + d\delta_{k3}$ where $d = \min_{\partial R \setminus S_0 \cup S_l} x_\alpha n_\alpha$ and consider the identity

$$\begin{aligned} 0 &= \frac{1}{\nu} \int_R y_k F_{i,k} (\nu \Delta F_i - \tilde{q}_{,i})\, d\mathscr{V} \\ &= -\frac{1}{\nu} \int_{\partial R} \tilde{q} y_k F_{i,k} n_i\, d\mathscr{S} + \int_{\partial R} y_k F_{i,k}(F_{i,j} - F_{j,i}) n_j\, d\mathscr{S} \\ &\quad + \frac{1}{2} \int_R F_{i,j}(F_{i,j} - F_{j,i})\, d\mathscr{V} - \frac{1}{2} \int_{\partial R} y_k n_k F_{i,j}(F_{i,j} - F_{j,i})\, d\mathscr{S}. \end{aligned} \tag{5.11}$$

Rewriting (5.11), we have

$$\frac{1}{2}\int_R F_{i,j}(F_{i,j}-F_{j,i})\,d\mathcal{V}=\frac{1}{\nu}\int_{S_0}\tilde{q}y_k(F_{i,k}n_i-F_{i,i}n_k)\,dx$$
$$-\int_{\partial R}y_k(n_jF_{i,k}-n_kF_{i,j})(F_{i,j}-F_{j,i})\,d\mathcal{S} \tag{5.12}$$
$$-\frac{1}{2}\int_{\partial R}y_kn_kF_{i,j}(F_{i,j}-F_{j,i})\,d\mathcal{S}.$$

It follows from Schwarz's inequality and the arithmetic-geometric mean inequality that

$$\frac{1}{2}\int_R F_{i,j}(F_{i,j}-F_{j,i})\,d\mathcal{V}$$
$$\leqq\left(\frac{a}{2}-\frac{1}{4}\right)\int_{\partial R}y_kn_k(F_{i,j}-F_{j,i})(F_{i,j}-F_{j,i})\,d\mathcal{S} \tag{5.13}$$
$$+\frac{1}{2a}\int_{S_0}\frac{y_sy_s}{y_ln_l}(n_jF_{i,k}-n_kF_{i,j})(n_jF_{i,k}-n_kF_{i,j})\,dx$$
$$+\frac{1}{2b}\int_{S_0}(y_\alpha F_{3,\alpha}-y_3F_{\alpha,\alpha})(y_\beta F_{3,\beta}-y_3F_{\beta,\beta})\,dx+\frac{b}{2\nu^2}\int_{S_0}\tilde{q}^2\,dx$$

for positive constants a and b, $a<\frac{1}{2}$. In Appendix III, we derive a bound for the last term of (5.13) of the form

$$\int_{S_0}\tilde{q}^2\,dx\leqq\frac{\nu^2\alpha_1}{4}\int_{\partial R}(F_{i,j}-F_{j,i})(F_{i,j}-F_{j,i})\,d\mathcal{S} \tag{5.14}$$

with an explicit constant α_1. Substituting (5.14) for the pressure term in (5.13), we obtain the result

$$\frac{1}{2}\int_R F_{i,j}(F_{i,j}-F_{j,i})\,d\mathcal{V}$$
$$\leqq\left[\left(\frac{a}{2}-\frac{1}{4}\right)d+\frac{b\alpha_1}{8}\right]\int_{\partial R}(F_{i,j}-F_{j,i})(F_{i,j}-F_{j,i})\,d\mathcal{S} \tag{5.15}$$
$$+\frac{1}{2a}\left(\frac{r_M^2+d^2}{d}\right)\int_{S_0}(n_jF_{i,k}-n_kF_{i,j})(n_jF_{i,k}-n_kF_{i,j})\,dx$$
$$+\frac{1}{2b}\int_{S_0}(y_\alpha F_{3,\alpha}-y_3F_{\alpha,\alpha})(y_\beta F_{3,\beta}-y_3F_{\beta,\beta})\,dx$$

where $r_M^2=\max_{S_0}x_\alpha x_\alpha$. We now choose a and b in such a way that $(a/2-\frac{1}{4})d+b\alpha_1/8=0$ and thus obtain from (5.15) a bound for $\int_R F_{i,j}(F_{i,j}-F_{j,i})\,d\mathcal{V}$ in terms of the given data and geometry. From (5.10) we then have the bound

$$\int_R F_{i,j}F_{i,j}\,d\mathcal{V}\leqq\frac{1}{a}\left(\frac{r_M^2+d^2}{d}\right)\int_{S_0}(n_jF_{i,k}-n_kF_{i,j})(n_jF_{i,k}-n_kF_{i,j})\,dx$$
$$+\frac{1}{b}\int_{S_0}(y_\alpha F_{3,\alpha}-y_3F_{\alpha,\alpha})(y_\beta F_{3,\beta}-y_3F_{\beta,\beta})\,dx \tag{5.16}$$
$$-\int_{S_0}(F_{3,\alpha}F_\alpha-F_{\alpha,\alpha}F_3)\,dx.$$

Using Schwarz's inequality and the arithmetic-geometric mean inequality leads to the result

$$\int_R F_{i,j}F_{i,j}\,d\mathcal{V} \leqq \gamma_1 \int_{S_0} F_iF_i\,dx + \gamma_2 \int_{S_0} F_{i,\alpha}F_{i,\alpha}\,dx \tag{5.17}$$

where the positive constants γ_1 and γ_2 can be explicitly computed and depend only on the geometry of the domain.

6. The Stokes flow problem for ζ_i. The purpose of this section is to briefly discuss the bounds for the flow associated with the velocity field ζ_i that are needed to complete our analysis. We first observe that the ζ_i satisfy

$$\begin{aligned} &\nu\Delta\zeta_i = \tilde{s}_{,i} \quad \text{in } R,\\ &\zeta_{i,i} = 0 \quad \text{in } R,\\ &\zeta_i = 0 \quad \text{on } \partial R\backslash S_l,\\ &\zeta_i = g_i - V\delta_{3i} \quad \text{on } S_l. \end{aligned} \tag{6.1}$$

This problem is just the Stokes problem (5.4) except that here the data are nonzero on S_l instead of S_0. Consequently, the bounds we need for this flow can be derived in much the same manner as those for (5.4) with some modifications.

Recall that we require estimates for $\max_z (\int_{S_z} \zeta_3^2\,dx)^{1/2}$, $\max_z (\int_{S_z} (\zeta_i\zeta_i)^2\,dx)^{1/4}$, and $\int_z^l \int_{S_\xi} (\xi - z)^n \zeta_{i,j}\zeta_{i,j}\,dx\,d\xi$ to make our decay inequality explicit. The first two quantities can be bounded using the methods of § 5.1 with the result that

$$\int_{S_z} \zeta_3^2\,dx \leqq \frac{1}{\sqrt{\lambda_1+\lambda_2}} \int_0^z \int_{S_\xi} \zeta_{i,j}\zeta_{i,j}\,dx\,d\xi \tag{6.2}$$

and

$$\int_{S_z} (\zeta_i\zeta_i)^2\,dx \leqq 8\Omega^{1/2}\left(\int_0^l \int_{S_\xi} \zeta_{i,j}\zeta_{i,j}\,dx\,dz\right)^2. \tag{6.3}$$

In addition, we observe that

$$J_n(z) = \int_z^l \int_{S_\xi} (\xi - z)^n \zeta_{i,j}\zeta_{i,j}\,dx\,d\xi \leqq l^n \int_0^l \int_S \zeta_{i,j}\zeta_{i,j}\,dx\,dz. \tag{6.4}$$

Thus, we again need to find an upper bound for the energy of a Stokes flow problem in terms of the data on S_l and geometry of the domain.

We will not go into the details of finding this bound but indicate where our previous analysis requires modification. Since the data are nonzero on S_l, we have upon integration and application of Schwarz's inequality

$$\begin{aligned} \int_R \zeta_{i,j}\zeta_{i,j}\,d\mathcal{V} &= \int_{S_l} \zeta_i(\zeta_{i,3} - \zeta_{3,i})\,dx + \int_{S_l} \zeta_i\zeta_{3,i}\,dx - \frac{1}{\nu}\int_{S_l} \zeta_3\tilde{s}\,dx\\ &\leqq \left(\int_{S_l} \zeta_i\zeta_i\,dx\right)^{1/2}\left(\int_{\partial R} (\zeta_{i,j} - \zeta_{j,i})(\zeta_{i,j} - \zeta_{j,i})\,d\mathcal{S}\right)^{1/2}\\ &\quad + \left(\int_{S_l} \zeta_3^2\,dx\right)^{1/2}\left(\int_{S_l} \tilde{s}^2\,dx\right)^{1/2} + \int_{S_l} [\zeta_\alpha\zeta_{3,\alpha} - \zeta_3\zeta_{\alpha,\alpha}]\,dx. \end{aligned}$$

The bound on $\int_{S_l} \tilde{s}^2\,dx$ can be found using the analysis of Appendix III with the obvious modification that $\partial G/\partial n = \tilde{s}$ on S_l and is zero elsewhere. To bound $\int_{\partial R} (\zeta_{i,j}-\zeta_{j,i})(\zeta_{i,j}-\zeta_{j,i})\,d\mathscr{S}$, we integrate, instead of (5.11), the identity

$$0 = \int_R [x_k - (l-d)\delta_{k3}]\zeta_{i,k}(\nu\Delta\zeta_i - \tilde{s}_{,i})\,d\mathscr{V}$$

where d is given following (5.10).

The final result is that we can compute explicit constants β_1, β_2 that depend only on the geometry of the domain such that

$$\int_0^l \int_S \zeta_{i,j}\zeta_{i,j}\,dx\,dz \leqq \beta_1 \int_{S_l} \zeta_i\zeta_i\,dx + \beta_2 \int_{S_l} \zeta_{i,\alpha}\zeta_{i,\alpha}\,dx. \tag{6.5}$$

We thus have bounds for the quantities that appear in M_1 and K as well as $J_n(z)$ $(n=2,3)$ in terms of the geometry and data of the problem.

For future use let us write

$$J_3(z) \leqq \kappa_3 l^3 Q_0 \tag{6.6}$$

where Q_0 is the data term in (6.5).

7. Discussion of results.

7.1. Comparison of entrance and fully-developed flows. In the last few sections, we have shown that provided certain conditions hold, $\Phi(z)$ satisfies an inequality of the form (4.2) where the constants K and M depend on v_s, ν, the boundary data, and the geometry. The data term N also depends on these quantities and in view of (6.4) and (6.5) is of order l^3. From (4.2), we can thus obtain the estimate (4.3) that we shall now use to compare u_i and $V\delta_{3i}$. Since $\omega_i = u_i - V\delta_{3i} = w_i + \zeta_i$, we see that

$$\begin{aligned}\left(\int_z^l \int_{S_\xi} (\xi-z)^3\omega_{i,j}\omega_{i,j}\,dx\,d\xi\right)^{1/2} &\leqq \left(\int_z^l \int_{S_\xi} (\xi-z)^3 w_{i,j}w_{i,j}\,dx\,d\xi\right)^{1/2} \\ &\quad + \left(\int_z^l \int_{S_\xi} (\xi-z)^3\zeta_{i,j}\zeta_{i,j}\,dx\,d\xi\right)^{1/2}.\end{aligned} \tag{7.1}$$

The first integral on the right of (7.1) is just a multiple of $\Phi(z)$ while the second is $\sqrt{J_3(z)}$. It then follows that

$$E(\omega_i) \equiv \frac{1}{3}\int_z^l \int_{S_\xi} (\xi-z)^3\omega_{i,j}\omega_{i,j}\,dx\,d\xi \leqq \left(1+\frac{\sqrt{\kappa_3}}{\sqrt{\kappa_2}}\right)\Phi(0)\,e^{-\kappa_1 z} + (\sqrt{\kappa_2}+\sqrt{\kappa_3})^2 l^3 Q_0 \tag{7.2}$$

where κ_2 and κ_3 are given by (4.3) and (6.6). Inequality (7.2) provides an upper bound for the weighted energy associated with the difference of velocities between the entrance flow and the fully developed flow measured over the portion of the pipe between S_z and S_l. If (7.2) is to be meaningful when l is large, then we need to assume that

$$\begin{aligned} Q_0 &\equiv \beta_1 \int_{S_l} \zeta_i\zeta_i\,dx + \beta_2 \int_{S_l} \zeta_{i,\alpha}\zeta_{i,\alpha}\,dx \\ &\equiv \beta_1 \int_{S_l} (g_i - V\delta_{3i})(g_i - V\delta_{3i})\,dx + \beta_2 \int_{S_l} (g_i - V\delta_{3i})_{,\alpha}(g_i - V\delta_{3i})_{,\alpha}\,dx \\ &\leqq \frac{\varepsilon^2}{l^3}\end{aligned} \tag{7.3}$$

for some $\varepsilon \ll 1$. In that case $E(\omega_i)$ will be bounded by a term that decays exponentially in z plus a term that is $O(\varepsilon^2)$.

7.2. Decay criteria. In the course of our analysis, we have imposed two restrictions on the flow to ensure that the first term in (7.2) decays. Let us examine these conditions more closely. We have from (4.30) and (5.9) the following conditions:

(1) $1-(v_s/2\nu\sqrt{\lambda_1})(1+2a_2)-k_2(1+a_1/2)\mathscr{H}^{1/2}>0$,

(2) $1-k_1v_s-k_2\mathscr{F}^{1/2}-k_3\mathscr{H}^{1/2}>0$

where the k_i, $\mathscr{H}$, and $\mathscr{F}$ are as defined in § 5 and a_1, a_2 are positive arbitrary constants, chosen so that condition (1) holds. Both of these conditions yield two pieces of information: (i) a restriction on ν and (ii) a restriction on the boundary data. From (1) we obtain

$$\beta_1\int_{S_l}\zeta_i\zeta_i\,dx+\beta_2\int_{S_l}\zeta_{i,\alpha}\zeta_{i,\alpha}\,dx\leqq\left\{\left[\frac{1}{k_2(1+a_1/2)}\right]\left[1-\frac{v_s}{2\nu\sqrt{\lambda_1}}(1+2a_2)\right]\right\}^2 \tag{7.4}$$

provided $1-(v_s/2\nu\sqrt{\lambda_1})(1+2a_2)>0$, or choosing $a_2=\varepsilon_2/2\ll 1$

$$\nu>\frac{v_s}{2\sqrt{\lambda_1}}(1+\varepsilon_2). \tag{7.5}$$

Condition (2) yields

$$\begin{aligned}&k_2\left(\gamma_1\int_{S_0}w_iw_i\,dx+\gamma_2\int_{S_0}w_{i,\alpha}w_{i,\alpha}\,dx\right)^{1/2}\\&\quad+k_3\left(\beta_1\int_{S_l}\zeta_i\zeta_i\,dx+\beta_2\int_{S_l}\zeta_{i,\alpha}\zeta_{i,\alpha}\,dx\right)^{1/2}\leqq 1-k_1v_s.\end{aligned} \tag{7.6}$$

We thus require $1-k_1v_s>0$ which, in view of our definition of k_1 translates into the viscosity restriction

$$\nu>\frac{v_s}{2\sqrt{\lambda_1}}\left[\frac{\sqrt{\lambda_1+\lambda_2}}{\sqrt{\lambda_2}}(1+\varepsilon_1)+\hat{c}\right]. \tag{7.7}$$

We are at liberty to choose $\hat{c}\ll 1$, but in any case if (7.7) is satisfied then so is (7.5). We thus see that condition (1) is actually contained in (2).

The two conditions (7.7) and (7.6) indicate that our inequality is valid only for flows with sufficiently large viscosity coefficients (or, with the proper definition, small Reynolds' numbers), and for flows whose data are suitably restricted.

7.3. Bounds for v_s and λ_2. To make our inequalities more explicit, we shall now demonstrate how an upper bound for v_s in terms of the geometry of the domain and the prescribed data can be obtained. We make use here of some results for the Saint-Venant torsion problem in a simply connected plane domain S. The torsion function Ψ satisfies

$$\Delta\Psi=-2 \quad \text{in } S, \qquad \Psi=0 \quad \text{on } \partial S.$$

In view of (2.11)-(2.13) and the relation $\tilde{p}_{,i}=-P\delta_{3i}$, we may express V in terms of the torsion function, the torsional rigidity $T\equiv 2\int_S\Psi\,dx$, and the net inflow Q. We have (see, e.g., [7])

$$V=\frac{2Q}{T}\Psi$$

and thus

$$v_s \leqq 2Q\frac{\Psi_M}{T}$$

where Ψ_M is the maximum value of the torsion function on S. We now need an upper bound on Ψ_M/T. For a simply connected domain, we have the isoperimetric inequality (see [12]),

$$2\pi\Psi_M^2 \leqq T.$$

Using this inequality, we see that

$$v_s \leqq Q\sqrt{\frac{2}{\pi T}}. \tag{7.8}$$

The desired upper bound can be obtained from (7.8) once we find a lower bound for the torsional rigidity. Such bounds for various domains are available in the literature or can be derived using monotony arguments [12], [15]. We mention here two results: (1) the inequality of Pólya and Szegö [15] for a star-shaped domain

$$T \geqq A^2R^{-1}$$

where A is the area of S and $R = \int_{\partial S}(xn_x + yn_y)^{-1}\,ds$, and (2) the bound obtained by Payne and Weinberger [14] for convex domains

$$T \geqq \frac{A^2}{2\pi}[1 - 2D^2(1-D^2)^{-1} - 4D^4(1-D^2)^{-2}\log D]$$

where $D^2 = 1 - 4\pi AL^{-2}$, A is the area of S, and L is the length of the perimeter.

Recall from § 3 that $\lambda_2 \geqq \lambda_1$. We can actually get a sharper bound for λ_2 in terms of the first two positive eigenvalues μ_1 $(=\lambda_1)$ and μ_2 of the fixed membrane problem using Weinstein's method [18] and the Payne-Rayner inequality [13]. We shall not go into the details of this derivation but merely record the lower bound that can be obtained. We find that

$$\lambda_2 \geqq \mu_1 + \left(\frac{\mu_2 - \mu_1}{A\mu_1}\right)4\pi$$

where A is the cross-sectional area.

7.4. Bound for $\Phi(0)$. Finally, we shall briefly indicate how an upper bound for $\Phi(0)$ can be derived. Using (4.9) with z replaced by zero, we may write $\Phi(0)$ as the sum of eleven integrals, each of which may be bounded in terms of $\Phi(0)$, $\int_0^l \int_S w_{i,j}w_{i,j}\,dx\,dz$, and $\int_0^l \int_S \zeta_{i,j}\zeta_{i,j}\,dx\,dz$. These bounds are generated in much the same way as those for the $\Phi_k(z)$ by making use of the appropriate integral inequalities. The major difference here is that it is necessary to use Hölder's inequality and Young's inequality in place of the Schwarz and arithmetic-geometric mean inequalities, respectively. We omit the derivation of these estimates since we assume by now the reader is familiar with the general procedure. The result is an inequality of the form

$$\mathcal{D}\Phi(0) \leqq \gamma_3 \int_0^l \int_S w_{i,j}w_{i,j}\,dx\,dz + \gamma_4 \int_0^l \int_S \zeta_{i,j}\zeta_{i,j}\,dx\,dz \tag{7.9}$$

where γ_3 and γ_4 are positive constants,

$$\mathcal{D} = 1 - k_1 v_s - k_3\mathcal{H}^{1/2} - \delta_1$$

and $0<\delta_1\ll 1$. Provided $\mathscr{D}>0$ and in view of the results of §§ 5 and 6, (7.9) yields an upper bound for $\Phi(0)$ in terms of the geometry and data. Note that we are guaranteed that $\mathscr{D}>0$ once we impose (5.9).

We conclude here by remarking that, in view of the results of this section, the constants in our estimates depend only on ν, the geometry of the domain, and the prescribed boundary data.

8. Concluding remarks. We note here that if we were to assume that v_i and the fully developed flow are identical, then we would recover Horgan and Wheeler's results with a slightly better decay constant. In fact, we would obtain instead of (4.2) the inequality

$$K\frac{d\Phi}{dz}+M\Phi\leqq 0$$

where $M=1-v_s/2\nu\sqrt{\lambda_1}$. Consequently, we may conclude that our methodology, although it bears much resemblance to that of the aforementioned authors, yields a slightly sharper decay rate.

We also observe that our estimate for $\Phi(z)$ can be used to obtain bounds on other integral quantities associated with the difference flow w_i. For example, we could find upper bounds for a weighted $\mathscr{L}_2$ integral of w_i or the ordinary $\mathscr{L}_2$ integral over a subdomain.

The case of a semi-infinite pipe is of some interest. In this case our results are still valid for $0\leqq z\leqq l$, where now we may think of Q_0 not as data but as the value of the combination of the "unknown" indicated integrals over that section. It becomes clear then from (7.2) and the arguments leading to (7.9) that if (5.9) is satisfied, and provided

$$\lim_{z\to\infty}\left[\beta_1\int_{S_z}\zeta_i\zeta_i\,dx+\beta_2\int_{S_z}\zeta_{i,\alpha}\zeta_{i,\alpha}\,dx\right]z^3=0,$$

we may conclude that

$$\int_z^\infty\int_{S_\xi}(\xi-z)^3w_{i,j}w_{i,j}\,dx\,d\xi\leqq\left[\int_0^\infty\int_{S_\xi}\xi^3w_{i,j}w_{i,j}\,dx\,d\xi\right]e^{-\kappa_1 z},$$

an inequality reminiscent of the exponential decay exhibited by the St. Venant Principle of classical elasticity.

Appendix I. In this Appendix, we derive the upper bounds given in (4.19)-(4.25) of the main text. We begin by considering

$$\Phi_1=-\int_z^l\int_{S_\xi}(\xi-z)^2w_iw_{i,3}\,dx\,d\xi.$$

Schwarz's inequality, (3.1), and the arithmetic-geometric mean inequality then yield

$$\Phi_1\leqq\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}\int_z^l\int_{S_\xi}(\xi-z)^2w_{i,j}w_{i,j}\,dx\,d\xi$$

from which (4.19) follows. For Φ_2, we see that

$$\begin{aligned}\Phi_2&=-\int_z^l\int_{S_\xi}(\xi-z)^2(\psi_\alpha w_{\alpha,3})_{,3}\,dx\,d\xi=2\int_z^l\int_{S_\xi}(\xi-z)\psi_\alpha w_{\alpha,3}\,dx\,d\xi\\&\leqq 2\left(\int_z^l\int_{S_\xi}(\xi-z)^2w_{\alpha,3}w_{\alpha,3}\,dx\,d\xi\right)^{1/2}\left(\int_z^l\int_{S_\xi}\psi_\alpha\psi_\alpha\,dx\,d\xi\right)^{1/2}.\end{aligned}\tag{I.1}$$

Now

$$\int_z^l \int_{S_\xi} \psi_\alpha \psi_\alpha \, dx \, d\xi = -2 \int_z^l \int_{S_\xi} (\xi - z) \psi_{\alpha,3} \psi_\alpha \, dx \, d\xi$$

$$\leqq 2 \left(\int_z^l \int_{S_\xi} \psi_\alpha \psi_\alpha \, dx \, d\xi \right)^{1/2} \left(\int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,3} \psi_{\alpha,3} \, dx \, d\xi \right)^{1/2}.$$

Thus, we find that

$$\left(\int_z^l \int_{S_\xi} \psi_\alpha \psi_\alpha \, dx \, d\xi \right)^{1/2} \leqq 2 \left(\int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,3} \psi_{\alpha,3} \, dx \, d\xi \right)^{1/2}$$

and, recalling that $\psi_{\alpha,3} = 0$ on ∂S_z, we have from (3.1) and (3.7) the bound

$$\left(\int_z^l \int_{S_\xi} \psi_\alpha \psi_\alpha \, dx \, d\xi \right)^{1/2} \leqq 2 \sqrt{\frac{C}{\lambda_1}} \left(\int_z^l \int_{S_\xi} (\xi - z)^2 w_{3,3}^2 \, dx \, d\xi \right)^{1/2}. \tag{I.2}$$

Substituting this result in (I.1) and applying the arithmetic-geometric mean inequality results in the estimate (4.20).

To bound Φ_4, we first integrate by parts and write it as the sum of three integrals:

$$\Phi_4 = \frac{1}{\nu} \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_\alpha w_{\alpha,j} w_j \, dx \, d\xi = \sum_{i=1}^{3} \Phi_4^i$$

$$= \frac{1}{\nu} \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,\beta} w_\beta w_\alpha \, dx \, d\xi - \frac{1}{\nu} \int_z^l \int_{S_\xi} (\xi - z)^2 \psi_{\alpha,3} w_3 w_\alpha \, dx \, d\xi$$

$$- \frac{2}{\nu} \int_z^l \int_{S_\xi} (\xi - z) \psi_\alpha w_\alpha w_3 \, dx \, d\xi.$$

Each of these three integrals can now be bounded using Schwarz's inequality, (3.1)-(3.3), (3.6), and the arithmetic-geometric mean inequality. We have

$$\Phi_4^1 \leqq \frac{1}{\nu} \int_z^l (\xi - z)^2 \left(\int_{S_\xi} (w_\alpha w_\alpha)^2 \, dx \right)^{1/4} \left(\int_{S_\xi} (w_\beta w_\beta)^2 \, dx \right)^{1/4} \left(\int_{S_\xi} \psi_{\alpha,\beta} \psi_{\alpha,\beta} \, dx \right)^{1/2} d\xi$$

$$\leqq \frac{\sqrt{C}}{\nu} \max_z \left(\int_{S_z} (w_\alpha w_\alpha)^2 \, dx \right)^{1/4}$$

$$\cdot \int_z^l (\xi - z)^2 \left(\int_{S_\xi} w_3^2 \, dx \right)^{1/2} \left[\frac{1}{2} \left(\int_{S_\xi} w_\alpha w_\alpha \, dx \right) \left(\int_{S_\xi} w_{\alpha,\beta} w_{\alpha,\beta} \, dx \right) \right]^{1/4} d\xi \tag{I.3}$$

$$\leqq \frac{1}{2\nu} \left(\frac{1}{2\lambda_1} \right)^{1/4} \sqrt{\frac{C}{\lambda_2}} \max_z \left(\int_{S_z} (w_\alpha w_\alpha)^2 \, dx \right)^{1/4}$$

$$\cdot \left\{ \int_z^l \int_{S_\xi} (\xi - z)^2 w_{3,\beta} w_{3,\beta} \, dx \, d\xi + \int_z^l \int_{S_\xi} (\xi - z)^2 w_{\alpha,\beta} w_{\alpha,\beta} \, dx \, d\xi \right\}.$$

An analogous process leads to the bound

$$\Phi_4^2 \leqq \frac{1}{2\nu} \left(\frac{1}{2\lambda_1} \right)^{1/4} \sqrt{\frac{C}{\lambda_2}} \max_z \left(\int_{S_z} (w_\alpha w_\alpha)^2 \, dx \right)^{1/4}$$

$$\cdot \left\{ \int_z^l \int_{S_\xi} (\xi - z)^2 w_{3,3}^2 \, dx \, d\xi + \int_z^l \int_{S_\xi} (\xi - z)^2 w_{3,\beta} w_{3,\beta} \, dx \, d\xi \right\}. \tag{I.4}$$

Finally, it follows from (3.3), (3.1), and (3.6) that

$$\int_{S_z}(\psi_\alpha\psi_\alpha)^2\,dx \leqq \frac{1}{2}\left(\int_{S_z}\psi_\alpha\psi_\alpha\,dx\right)\left(\int_{S_z}\psi_{\alpha,\beta}\psi_{\alpha,\beta}\,dx\right) \leqq \frac{1}{2\lambda_1}\left(\int_{S_z}\psi_{\alpha,\beta}\psi_{\alpha,\beta}\,dx\right)^2$$
$$\leqq \frac{C^2}{2\lambda_1}\left(\int_{S_z} w_3^2\,dx\right)^2$$

and therefore

$$\Phi_4^3 \leqq \frac{2}{\nu}\int_z^l (\xi - z)\left(\int_{S_\xi}(\psi_\alpha\psi_\alpha)^2\,dx\right)^{1/4}\left(\int_{S_\xi}(w_\alpha w_\alpha)^2\,dx\right)^{1/4}\left(\int_{S_\xi} w_3^2\,dx\right)^{1/2} d\xi$$
$$\leqq \frac{2}{\nu}\left(\frac{C^2}{2\lambda_1}\right)^{1/4}\max_z\left(\int_{S_z}(w_\alpha w_\alpha)^2\,dx\right)^{1/4}\int_z^l\int_{S_\xi}(\xi - z)w_3^2\,dx\,d\xi.$$

Now, since $\int_z^l\int_{S_\xi}(\xi - z)w_3^2\,dx\,d\xi = -\int_z^l\int_{S_\xi}(\xi - z)w_3 w_{3,3}\,dx\,d\xi$, we find after application of (3.2) and the arithmetic-geometric mean inequality that

$$\Phi_4^3 \leqq \frac{1}{\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\sqrt{\frac{C}{\lambda_2}}\max_z\left(\int_{S_z}(w_\alpha w_\alpha)^2\,dx\right)^{1/4}$$
$$\cdot\left\{\int_z^l\int_{S_\xi}(\xi - z)^2 w_{3,\beta}w_{3,\beta}\,dx\,d\xi + \int_z^l\int_{S_\xi}(\xi - z)^2 w_{3,3}^2\,dx\,d\xi\right\}. \tag{I.5}$$

Combining (I.3)-(I.5), we then obtain (4.21).

Turning now to Φ_6, we write

$$\Phi_6 = \frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi - z)^2\psi_\alpha w_j v_{\alpha,j}\,dx\,d\xi = -\frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi - z)^2\psi_{\alpha,\beta}w_\beta\zeta_\alpha\,dx\,d\xi$$
$$-\frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi - z)^2\psi_{\alpha,3}w_3\zeta_\alpha\,dx\,d\xi - \frac{2}{\nu}\int_z^l\int_{S_\xi}(\xi - z)\psi_\alpha w_3\zeta_\alpha\,dx\,d\xi$$

where $\zeta_i = v_i - V\delta_{3i}$.

Again, the appropriate inequalities applied to each of the three integrals in this expression for Φ_6 will lead to the bound

$$\Phi_6 \leqq \frac{1}{2\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\sqrt{\frac{C}{\lambda_2}}\max_z\left(\int_{S_z}(\zeta_\alpha\zeta_\alpha)^2\right)^{1/4}$$
$$\cdot\left[\int_z^l\int_{S_\xi}(\xi - z)^2 w_{3,\beta}w_{3,\beta}\,dx\,d\xi + \int_z^l\int_{S_\xi}(\xi - z)^2 w_{\alpha,\beta}w_{\alpha,\beta}\,dx\,d\xi\right]$$
$$+\frac{5}{2\nu}\left(\frac{1}{2\lambda_2}\right)^{1/4}\sqrt{\frac{C}{\lambda_1}}\max_z\left(\int_{S_z}(\zeta_\alpha\zeta_\alpha)^2\right)^{1/4}$$
$$\cdot\left[\int_z^l\int_{S_\xi}(\xi - z)^2 w_{3,\beta}w_{3,\beta}\,dx\,d\xi + \int_z^l\int_{S_\xi}(\xi - z)^2 w_{3,3}^2\,dx\,d\xi\right]$$

from which (4.22) easily follows.

Consider now

$$\Phi_7 = \frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi - z)^2\psi_\alpha v_j v_{\alpha,j}\,dx\,d\xi$$
$$= \frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi - z)^2\psi_\alpha\zeta_j\zeta_{\alpha,j}\,dx\,d\xi + \frac{1}{\nu}\int_z^l\int_{S_\xi}(\xi - z)^2\psi_\alpha V\zeta_{\alpha,3}\,dx\,d\xi$$
$$\leqq \frac{1}{\nu}\int_z^l(\xi - z)^2\left(\int_{S_\xi}(\psi_\alpha\psi_\alpha)^2\,dx\right)^{1/4}\left(\int_{S_\xi}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}\left(\int_{S_\xi}(\zeta_{\alpha,j}\zeta_{\alpha,j})^2\,dx\right)^{1/2} d\xi$$

$$(I.6)\qquad +\frac{v_s}{\nu}\int_z^l (\xi-z)^2\left(\int_{S_\xi}(\psi_\alpha\psi_\alpha)^2\,dx\right)^{1/2}\left(\int_{S_\xi}\zeta_{\alpha,3}\zeta_{\alpha,3}\,dx\right)^{1/2}d\xi$$

$$\leqq\frac{1}{2\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\sqrt{\frac{C}{\lambda_2}}\max_z\left(\int_{S_z}(\zeta_i\zeta_i)^2\,dx\right)^{1/4}$$

$$\cdot\left[\int_z^l\int_{S_\xi}(\xi-z)^2w_{3,\beta}w_{3,\beta}\,dx\,d\xi+\int_z^l\int_{S_\xi}(\xi-z)^2\zeta_{\alpha,j}\zeta_{\alpha,j}\,dx\,d\xi\right]$$

$$+\frac{v_s}{2\nu\sqrt{\lambda_1}}\sqrt{\frac{C}{\lambda_2}}\left[\int_z^l\int_{S_\xi}(\xi-z)^2w_{3,\beta}w_{3,\beta}\,dx\,d\xi+\int_z^l\int_{S_\xi}(\xi-z)^2\zeta_{\alpha,3}\zeta_{\alpha,3}\,dx\,d\xi\right].$$

Inequality (4.23) then follows directly from (I.6).

Finally, bounds for Φ_8 and Φ_{11} are easily derivable using the same procedure. We find

$$\Phi_8=\frac{1}{2\nu}\int_z^l\int_{S_\xi}(\xi-z)^2w_iw_iv_3\,dx\,d\xi$$

$$=\frac{1}{2\nu}\int_z^l\int_{S_\xi}(\xi-z)^2w_iw_i\zeta_3\,dx\,d\xi+\frac{1}{2\nu}\int_z^l\int_{S_\xi}(\xi-z)^2w_iw_iV\,dx\,d\xi$$

$$\leqq\left[\frac{1}{2\nu}\frac{1}{\sqrt{2\lambda_1}}\max_z\left(\int_{S_z}\zeta_3^2\,dx\right)^{1/2}+\frac{v_s}{2\nu\lambda_1}\right]\int_z^l\int_{S_\xi}(\xi-z)^2w_{i,\beta}w_{i,\beta}\,dx\,d\xi$$

and

$$\Phi_{11}=\frac{1}{2\nu}\int_z^l\int_{S_\xi}(\xi-z)^2w_3w_iw_i\,dx\,d\xi$$

$$\leqq\frac{1}{2\nu\sqrt{2\lambda_1}}\max_z\left(\int_{S_z}w_3^2\,dx\right)^{1/2}\int_z^l\int_{S_\xi}(\xi-z)^2w_{i,\beta}w_{i,\beta}\,dx\,d\xi.$$

Appendix II. We will briefly indicate here how the inequality (5.7) is obtained from (5.5) and (5.6). The basic idea is to bound each of the integrals I_K in (5.6) in terms of I, $\int_0^l\int_S w_{i,j}w_{i,j}\,dx\,dz$, $\int_0^l\int_S F_{i,j}F_{i,j}\,dx\,dz$, and $\int_0^l\int_S\zeta_{i,j}\zeta_{i,j}\,dx\,dz$. Such bounds are derived by appropriately utilizing Schwarz's inequality, (3.1)–(3.3), and the arithmetic-geometric mean inequality. Since this is a somewhat tedious task, we will illustrate the methodology for only one integral and summarize the bounds for the other eight.

Let us consider I_3. We have

$$I_3=\frac{1}{\nu}\int_0^l\int_S(w_i-F_i)_{,j}w_jF_i\,dx\,dz$$

$$\leqq\frac{1}{\nu}\max_z\left(\int_S(F_iF_i)^2\,dx\right)^{1/4}\left(\int_0^l\int_S(w_i-F_i)_{,j}(w_i-F_i)_{,j}\,dx\,dz\right)^{1/2}$$

$$\cdot\left(\int_0^l\int_S(w_jw_j)^2\,dx\,dz\right)^{1/4}$$

$$\leqq\frac{1}{2\nu}\left(\frac{1}{2\lambda_1}\right)^{1/4}\max_z\left(\int_S(F_iF_i)^2\,dx\right)^{1/4}$$

$$\cdot\left[a_1\int_0^l\int_S(w_i-F_i)_{,j}(w_i-F_i)_{,j}\,dx\,dz+\frac{1}{a_1}\int_0^l\int_S w_{i,\beta}w_{i,\beta}\,dx\,dz\right]$$

after applying Schwarz's inequality twice and then using the Sobolev inequality (3.3) and arithmetic-geometric mean inequality.

We observe that the bounds for I_2, I_7, I_8, and I_9 can be combined in such a way that the constants arising from the arithmetic-geometric mean inequality can be chosen so as to optimize the coefficient of $\int_0^l \int_S (w_i - F_i)_{,j}(w_i - F_i)_{,j}\, dx\, dz$. In fact, we find that

$$I_2 + I_7 + I_8 + I_9 \leqq \frac{v_s}{2\nu\sqrt{\lambda_1}} \sqrt{\frac{\lambda_1+\lambda_2}{\lambda_2}} (1+\varepsilon_1) \int_0^l \int_S (w_i - F_i)_{,j}(w_i - F_i)_{,j}\, dx\, dz$$
$$+ \frac{v_s}{\varepsilon_1 \nu\sqrt{\lambda_1}} \sqrt{\frac{\lambda_1+\lambda_2}{\lambda_2}} \int_0^l \int_S \zeta_{i,j}\zeta_{i,j}\, dx\, dz$$

where $0 < \varepsilon_1 \ll 1$.

Combining our bounds, we have

$$\begin{aligned} I = &\int_0^l \int_S (w_i - F_i)_{,j}(w_i - F_i)_{,j}\, dx\, dz \\ \leqq &\left\{ \frac{v_s}{2\nu\sqrt{\lambda_1}} \sqrt{\frac{\lambda_1+\lambda_2}{\lambda_2}} (1+\varepsilon_1) + \hat{c} + \frac{b_2}{2\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \max_z \left(\int_S (F_i F_i)^2\, dx \right)^{1/4} \right. \\ &\left. + \frac{1}{\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \left(1 + a_1 + \frac{b_1}{2}\right) \max_z \left(\int_S (\zeta_i \zeta_i)^2\, dx \right)^{1/4} \right\} I \\ &+ \frac{1}{2b_2\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \max_z \left(\int_S (F_i F_i)^2\, dx \right)^{1/4} \int_0^l \int_S w_{i,j} w_{i,j}\, dx\, dz \\ &+ \left\{ \frac{2v_s}{\hat{c}\nu\sqrt{\lambda_1}} + \frac{1}{a_1\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \max_z \left(\int_S (\zeta_i \zeta_i)^2\, dx \right)^{1/4} \right\} \int_0^l \int_S F_{i,j} F_{i,j}\, dx\, dz \\ &+ \left\{ \frac{v_s}{\nu\varepsilon_1\sqrt{\lambda_1}} \sqrt{\frac{\lambda_1+\lambda_2}{\lambda_2}} + \frac{1}{2b_1\nu} \left(\frac{1}{\lambda_1}\right)^{1/4} \max_z \left(\int_S (\zeta_i \zeta_i)^2\, dx \right)^{1/4} \right\} \int_0^l \int_S \zeta_{i,j}\zeta_{i,j}\, dx\, dz \end{aligned}$$

where a_i, b_i $(i = 1, 2)$ and $\hat{c}$ are arbitrary positive constants. If we now make use of the identity (5.5), choose $b_2 = 1$, and rearrange the preceding inequality, we obtain

$$\begin{aligned} &\left\{ 1 - k_1 v_s - \frac{1}{\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \left(1 + a_1 + \frac{b_1}{2}\right) \mathcal{M}_1 - \frac{1}{\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \mathcal{M}_2 \right\} \int_0^l \int_S w_{i,j} w_{i,j}\, dx\, dz \\ &\quad \leqq \left\{ \frac{v_s}{\varepsilon\nu\sqrt{\lambda_1}} \sqrt{\frac{\lambda_1+\lambda_2}{\lambda_2}} + \frac{1}{2b_1\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \mathcal{M}_1 \right\} \int_0^l \int_S \zeta_{i,j}\zeta_{i,j}\, dx\, dz \\ &\quad + \left\{ 1 - k_1 v_s + \frac{2v_s}{\hat{c}\nu\sqrt{\lambda_1}} - \frac{1}{\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \left(1 + a_1 + \frac{b_1}{2} - \frac{1}{a_1}\right) \mathcal{M}_1 - \frac{1}{2\nu} \left(\frac{1}{2\lambda_1}\right)^{1/4} \mathcal{M}_2 \right\} \\ &\quad \cdot \int_0^l \int_S F_{i,j} F_{i,j}\, dx\, dz \end{aligned} \tag{II.1}$$

where

$$k_1 = \frac{1}{2\nu\sqrt{\lambda_1}} \sqrt{\frac{\lambda_1+\lambda_2}{\lambda_2}} (1+\varepsilon_1)$$

and

$$\mathcal{M}_1 = \max_z \left(\int_S (\zeta_i \zeta_i)^2\, dx \right)^{1/4}, \qquad \mathcal{M}_2 = \max_z \left(\int_S (F_i F_i)^2\, dx \right)^{1/4}.$$

It now remains to find bounds for $\mathcal{M}_1$ and $\mathcal{M}_2$. We observe that since $\zeta_i = 0$ on S_0

$$\begin{aligned}\int_S (\zeta_i\zeta_i)^2\,dx &= 4\int_0^z\int_{S_\xi} \zeta_i\zeta_i\zeta_j\zeta_{j,3}\,dx\,d\xi \\ &\leqq 4\left(\int_0^z\int_{S_\xi}(\zeta_i\zeta_i)^3\,dx\,d\xi\right)^{1/2}\left(\int_0^z\int_{S_\xi}\zeta_{j,3}\zeta_{j,3}\,dx\,d\xi\right)^{1/2} \\ &\leqq 8\Omega^{1/2}\left(\int_0^l\int_{S_\xi}\zeta_{i,j}\zeta_{i,j}\,dx\,dz\right)^2\end{aligned} \tag{II.2}$$

where we have used (3.4) to obtain (II.2). Consequently,

$$\mathcal{M}_1 \leqq 2^{3/4}\,\Omega^{1/8}\left(\int_0^l\int_S \zeta_{i,j}\zeta_{i,j}\,dx\,dz\right)^{1/2}. \tag{II.3}$$

Similarly, since $F_i = 0$ on S_l, we find that

$$\mathcal{M}_2 \leqq 2^{3/4}\,\Omega^{1/8}\left(\int_0^l\int_S F_{i,j}F_{i,j}\,dx\,dz\right)^{1/2}. \tag{II.4}$$

Substituting (II.3) and (II.4) into (II.1) and choosing $b_1 = 2a_1 = \delta$ for $0 < \delta \ll 1$, we obtain the results (5.7) and (5.8).

Appendix III. To derive the bound (5.15) for the pressure term in the Stokes flow problem, we first observe that $\tilde{q}$ is defined only up to an arbitrary constant and then fix it by requiring $\int_{S_0}\tilde{q}\,dx = 0$. Let us now introduce an auxiliary function G such that

$$\begin{aligned}&\Delta G = 0 \quad \text{in } R = S\times[0,l], \\ &\frac{\partial G}{\partial n} = 0 \quad \text{on } \partial R\backslash S_0, \\ &\frac{\partial G}{\partial n} = \tilde{q} \quad \text{on } S_0.\end{aligned} \tag{III.1}$$

We have

$$\int_{S_0}\tilde{q}^2\,dx = \int_{\partial R}\tilde{q}\frac{\partial G}{\partial n}\,d\mathcal{S} = \int_R \tilde{q}_{,i}G_{,i}\,d\mathcal{V} = \nu\int_R G_{,i}(F_{i,j}-F_{j,i})_{,j}\,d\mathcal{V}$$

and thus

$$\int_{S_0}\tilde{q}^2\,dx = \frac{\nu}{2}\int_{\partial R}(G_{,i}n_j - G_{,j}n_i)(F_{i,j}-F_{j,i})\,d\mathcal{S}.$$

An application of Schwarz's inequality yields the bound

$$\begin{aligned}\int_{S_0}\tilde{q}^2\,dx \leqq \frac{\nu}{2}&\left(\int_{\partial R}(F_{i,j}-F_{j,i})(F_{i,j}-F_{j,i})\,d\mathcal{S}\right)^{1/2} \\ &\cdot\left(\int_{\partial R}(G_{,i}n_j - G_{,j}n_i)(G_{,i}n_j - G_{,j}n_i)\,d\mathcal{S}\right)^{1/2}.\end{aligned} \tag{III.2}$$

We now need to bound the second integral on the right in terms of $\int_{S_0}(\partial G/\partial n)^2\,dx$.

For a region whose boundary is star-shaped with respect to the origin (which we take at the center of the section $z = l/2$), we have the following two identities:

$$\begin{aligned} 0 &= \int_0^l \int_S \left(\xi - \frac{l}{2}\right) G_{,\xi} \Delta G \, dx \, d\xi \\ &= \frac{1}{2}\int_0^l \int_S [G_{,\alpha}G_{,\alpha} - G_{,3}^2] \, dx \, d\xi - \frac{l}{4}\int_{S_0 \cup S_l} G_{,\alpha}G_{,\alpha} \, dx + \frac{l}{4}\int_{S_0} \tilde{q}^2 \, dx, \end{aligned} \tag{III.3}$$

$$\begin{aligned} 0 &= \int_0^l \int_S x_\alpha G_{,\alpha} \Delta G \, dx \, d\xi \\ &= \int_0^l \int_S G_{,3}^2 \, dx \, d\xi - \frac{1}{2}\int_{\partial S \times [0,l]} x_\alpha n_\alpha G_{,j} G_{,j} \, d\mathcal{S} + \int_{S_0} x_\alpha G_{,\alpha} \tilde{q} \, dx. \end{aligned} \tag{III.4}$$

Upon rewriting (III.3) and (III.4), we see that

(III.5)

$$\int_{S_0 \cup S_l} G_{,\alpha}G_{,\alpha} \, dx = \int_{S_0 \cup S_l} |\text{grad}_s \, G|^2 \, dx = \frac{2}{l}\int_0^l \int_S [G_{,\alpha}G_{,\alpha} - G_{,3}^2] \, dx \, d\xi + \int_{S_0} \tilde{q}^2 \, dx$$

and

$$\frac{1}{2}\int_{\partial S \times [0,l]} x_\alpha n_\alpha |\text{grad}_s \, G|^2 \, d\mathcal{S} = \int_{S_0} x_\alpha G_{,\alpha} \tilde{q} \, dx + \int_0^l \int_S G_{,3}^2 \, dx \, d\xi \tag{III.6}$$

where $|\text{grad}_s \, G|^2$ denotes the tangential derivatives, i.e.,

$$|\text{grad}_s \, G|^2 = \tfrac{1}{2}(n_k G_{,j} - n_j G_{,k})(n_k G_{,j} - n_j G_{,k}).$$

An application of Schwarz's inequality and the arithmetic-geometric mean inequality in (III.6) leads to the inequality

$$\int_{\partial S \times [0,l]} |\text{grad}_s \, G|^2 \, d\mathcal{S} \leqq \sigma \frac{r_M^2}{h}\int_{S_0} \tilde{q}^2 \, dx + \frac{1}{\sigma h}\int_{S_0} G_{,\alpha}G_{,\alpha} \, dx + \frac{2}{h}\int_0^l \int_S G_{,3}^2 \, dx \, d\xi \tag{III.7}$$

where $r_M^2 = \max_{S_0} x_\alpha x_\alpha$, $h = \min_{\partial S} x_\alpha n_\alpha$, and σ is a positive constant. If we now add (III.5) and (III.7) and choose $\sigma = 1/r_M$, we obtain

$$\int_{\partial R} |\text{grad}_s \, G|^2 \, d\mathcal{S} \leqq \left(1 + \frac{2r_M}{h}\right)\int_{S_0} \tilde{q}^2 \, dx + K \int_0^l \int_S G_{,j}G_{,j} \, dx \, d\xi \tag{III.8}$$

where

$$K = \max\left[\frac{2}{l}\left(1 + \frac{r_M}{h}\right), \frac{2}{h} - \frac{2}{l}\left(1 + \frac{r_M}{h}\right)\right].$$

Let us now consider $\int_0^l \int_S G_{,j}G_{,j} \, dx \, d\xi = J$. Since

$$J = \int_{S_0} G\tilde{q} \, dx \leqq \left(\int_{S_0} G^2 \, dx\right)^{1/2}\left(\int_{S_0} \tilde{q}^2 \, dx\right)^{1/2}$$

it follows that

$$J^2 \leqq \frac{1}{\mu_2}\left(\int_{S_0} G_{,\alpha}G_{,\alpha} \, dx\right)\left(\int_{S_0} \tilde{q}^2 \, dx\right)$$

where μ_2 is the first positive eigenvalue of the free membrane problem. In view of (III.5), we have

$$\mu_2 J^2 \leqq M^2 + \frac{2}{l} MJ$$

where $M = \int_{S_0} \tilde{q}^2 \, dx$. Solving this inequality, we conclude that

$$\text{(III.9)} \qquad \int_0^l \int_S G_{,j} G_{,j} \, dx \, d\xi \leqq \frac{1}{\mu_2 l} (1 + \sqrt{1 + \mu_2^2 l^2}) \int_{S_0} \tilde{q}^2 \, dx.$$

Inserting this inequality into (III.8), we then obtain the bound

$$\int_{\partial R} |\text{grad}_S \, G|^2 \, d\mathcal{S} \leqq \alpha_1 \int_{S_0} \tilde{q}^2 \, dx$$

where

$$\alpha_1 = \left(1 + \frac{2r_M}{h}\right) + \frac{K}{\mu_2 l} (1 + \sqrt{1 + \mu_2^2 l^2}).$$

From (III.2), it follows that

$$\int_{S_0} \tilde{q}^2 \, dx \leqq \frac{\alpha_1 \nu^2}{4} \int_{\partial R} (F_{i,j} - F_{j,i})(F_{i,j} - F_{j,i}) \, d\mathcal{S}.$$

REFERENCES

[1] C. J. AMICK, *Steady solutions of the Navier-Stokes equations in unbounded channels and pipes*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 4 (1977), pp. 473-513.

[2] ———, *Properties of steady Navier-Stokes solutions for certain unbounded channels and pipes*, Nonlinear Anal., 2 (1978), pp. 689-720.

[3] 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 Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 5-359.

[4] C. O. HORGAN, *Plane steady flows and energy estimates for the Navier-Stokes equations*, Arch. Rational Mech. Anal., 68 (1978), pp. 359-381.

[5] C. O. HORGAN AND J. K. KNOWLES, *Recent developments concerning Saint-Venant's Principle*, Adv. in Appl. Mech., 23 (1983), pp. 179-269.

[6] C. O. HORGAN AND L. E. PAYNE, *On inequalities of Korn, Friedrichs and Babuška-Aziz*, Arch. Rational Mech. Anal., 82 (1983), pp. 165-179.

[7] C. O. HORGAN AND L. T. WHEELER, *Spatial decay estimates for the Navier-Stokes equations with applications to the problem of entry flow*, SIAM J. Appl. Math., 35 (1978), pp. 97-116.

[8] J. K. KNOWLES, *On Saint-Venant's Principle in the two-dimensional linear theory of elasticity*, Arch. Rational Mech. Anal., 21 (1966), pp. 1-22.

[9] H. A. LEVINE, *An estimate for the best constant in a Sobolev inequality involving three integral norms*, Ann. Mat. Pura Appl. (4), 124 (1980), pp. 181-197.

[10] O. A. OLEINIK, *Applications of the energy estimates analogous to Saint-Venant's Principle to problems of elasticity and hydrodynamics*, Lecture Notes in Physics 90, Springer-Verlag, Berlin, New York, 1979, pp. 422-432.

[11] L. E. PAYNE, *Uniqueness criteria for steady state solutions of the Navier-Stokes equations*, Simpos. Internaz. Appl. Anal. Fis. Mat. (Cagliari-Sassari, 1964), Edizioni Cremonese, Rome, 1965, pp. 130-153.

[12] ———, *Isoperimetric inequalities and their applications*, SIAM Rev., 9 (1967), pp. 453-488.

[13] L. E. PAYNE AND M. E. RAYNER, *An isoperimetric inequality for the first eigenfunction in the fixed membrane problem*, Z. Angew. Math. Phys., 23 (1972), pp. 13-15.

[14] L. E. PAYNE AND H. F. WEINBERGER, *Some isoperimetric inequalities for membrane frequencies and torsional rigidity*. J. Math. Anal. Appl., 2 (1961), pp. 210-216.

[15] G. PÓLYA AND G. SZEGÖ, *Isoperimetric Inequalities in Mathematical Physics*, Ann. Math. Stud., 27, Princeton University Press, Princeton, NJ, 1951.

[16] G. TALENTI, *Best constant in Sobolev inequality*, Ann. Mat. Pura. Appl. (4), 110 (1976), pp. 353-372.

[17] R. A. TOUPIN, *Saint-Venant's Principle*, Arch. Rational Mech. Anal., 18 (1965), pp. 83-96.

[18] A. WEINSTEIN AND W. STENGER, *Methods of Intermediate Problems for Eigenvalues: Theory and Ramifications*, Academic Press, New York, 1972.

[19] G. A. YOSIFIAN, *An analog of Saint-Venant's Principle and the uniqueness of the solutions of the first boundary value problem for Stokes' system in domains with noncompact boundaries*, Dokl. Akad. Nauk. SSSR, 242 (1978), pp. 36-39. (In Russian.) Soviet Math. Dokl., 19 (1978), pp. 1048-1052. (In English.)

[20] ———, *Saint-Venant's Principle for the flow of a viscous incompressible liquid*, Uspekhi Mat. Nauk., 34 (1979), pp. 191-192. (In Russian.) Russian Math. Surveys, 34 (1979), pp. 166-167. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 816-844, July 1989

004

FINITE-DIMENSIONAL ATTRACTORS ASSOCIATED WITH PARTLY DISSIPATIVE REACTION-DIFFUSION SYSTEMS*

MARTINE MARION†

Résumé. L'objet de cet article est d'étudier le comportement quand le temps tend vers l'infini des solutions de certains systèmes de réaction-diffusion partiellement dissipatifs. On considère deux types de problèmes: des systèmes avec des nonlinéarités polynomiales et des systèmes possédant une région positivement invariante. L'article démontre que le comportement à t infini peut être décrit par un attracteur universel, et des majorations des dimensions Hausdorff et fractale de cet attracteur sont établis. Ces résultats sont appliqués à plusieurs systèmes classiques issus de la biologie, de la physique et de la chimie.

Abstract. The long-time behavior of the solutions of some partly dissipative reaction-diffusion systems is studied. Two types of problems are considered: systems with a polynomial growth nonlinearity, and systems admitting a positively invariant region. It is shown that the long-time behavior can be described by a universal attractor, and bounds of the Hausdorff and fractal dimensions of this attractor are derived. The results are applied to several classical systems borrowed from mathematical biology, physics and chemistry.

Key words. attractors, reaction-diffusion equations, partly dissipative systems, fractal dimension

AMS(MOS) subject classifications. 35B40, 35K57, 35Q20

0. Introduction. It is now well known that many parabolic dissipative evolutionary equations possess a universal attractor that has finite Hausdorff and fractal dimensions. All trajectories converge to this attractor as time goes to infinity so that the long-time behavior of solutions depends actually on a finite number of degrees of freedom (see Temam [26] for an extensive review on the subject). In particular, in the case of reaction-diffusion equations, such questions have been investigated by Babin and Vishik [1], [2], Kopell and Ruelle [15], and Marion [18]. However, in these works, a strong dissipation assumption is required; namely, all diffusion coefficients are assumed to be positive.

The aim of this article is to derive the existence of finite-dimensional attractors for reaction-diffusion systems with vanishing diffusion coefficients. The mathematical study differs from the ones in [1], [2], [15], [18], mainly for two reasons:

(1) The semigroup associated with the problem is no longer compact.

(2) The methods used in [1], [2], [15], [18] to prove the finite dimensionality fail; the upper bound on the dimension of the attractor derived there goes to $+\infty$ as the diffusion tends to zero.

We will investigate two types of systems: those with a polynomial growth nonlinearity, and those admitting a positively invariant region.

We first study systems with a polynomial growth nonlinearity. A typical sample of systems we consider is

$$\frac{\partial u}{\partial t} - d\Delta u + h(u) + \sigma v = 0, \qquad \frac{\partial v}{\partial t} + \delta v + \gamma u = 0, \tag{0.1}$$

where $\delta > 0$, $\sigma, \gamma \in \mathbb{R}$, and h is a polynomial of odd degree with a positive leading coefficient. The precise assumptions on the equations are stated in § 1. In § 2, we derive the existence of a universal attractor by applying a general criterion of existence of

* Received by the editors December 1, 1987; accepted for publication July 1, 1988.

† Laboratoire d'Analyse Numérique, Bâtiment 425, Université Paris-Sud, 91405 Orsay Cedex, France.

attractors used by Ghidaglia and Temam [10] for abstract nonlinear wave equations (see also the review of Hale [27] and the presentation in Temam [26]). We then prove an important regularity property of the attractor. Section 3 contains the estimates of the Hausdorff and fractal dimensions of the attractor; these estimates rely on the general method of Constantin, Foias and Temam [6] and on generalizations of the Sobolev-Lieb-Thirring inequalities proved in Ghidaglia, Marion, and Temam [11]. We conclude the section by applying our results to problem (0.1).

We then consider systems of reaction-diffusion equations admitting a positively invariant region as described by Chueh, Conley, and Smoller [3]. Section 4 contains the description of the equations; we assume in particular that the equations take the form

$$\frac{\partial u}{\partial t} - D\Delta u + f(x, u, v) = 0, \qquad \frac{\partial v}{\partial t} + G(x, u)v + g(x, u) = 0,$$

where u (respectively, v) takes its values in $\mathbb{R}^{m_1}$ (respectively, $\mathbb{R}^{m_2}$) and $G(x, u)$ is a square matrix of order m_2 with definite positive symmetric part. The existence of a universal attractor is proved in § 5. We also derive there the estimates of the Hausdorff and fractal dimensions of the attractor. In § 6 we apply our results to several classical systems, namely the nerve equations (Hodgkin-Huxley equations and FitzHugh-Nagumo equations), some equations related to solid combustion, and the Feld-Noyes equations arising in chemical kinetics.

The results presented here were announced in [28].

A few words about notation follow. Let Ω be an open bounded set of $\mathbb{R}^n$ with boundary Γ. For $p \in [1, +\infty]$, we denote by $L^p(\Omega)$ the space of measurable scalar functions on Ω for which

$$|u|_{L^p(\Omega)} = \left(\int_\Omega |u(x)|^p \, dx \right)^{1/p} < +\infty \quad \text{for } 1 \leqq p < \infty,$$

$$|u|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x \in \Omega} |u(x)| < +\infty \quad \text{for } p = +\infty.$$

We denote by $H^k(\Omega)$ the Sobolev space of scalar functions that are in $L^2(\Omega)$ together with their weak derivatives of order less than or equal to $k \in \mathbb{N}^*$. $H_0^1(\Omega)$ is the Hilbert subspace of $H^1(\Omega)$ made of functions vanishing on Γ.

We will also consider vector-valued functions and use the notation $\mathbb{L}^2(\Omega) = (L^2(\Omega))^m$, $\mathbb{H}^k(\Omega) = (H^k(\Omega))^m$, $\mathbb{H}_0^1(\Omega) = (H_0^1(\Omega))^m$. We denote by $(\cdot, \cdot)$ the scalar product on $\mathbb{L}^2(\Omega)$ and we equip $\mathbb{H}_0^1(\Omega)$ with the norm

$$\|u\| = \left(\sum_{i=1}^n \left| \frac{\partial u}{\partial x_1} \right|^2 \right)^{1/2}.$$

Let I be a bounded interval of $\mathbb{R}$ and let X be a Banach space. We denote by $L^p(I; X)$, $1 \leqq p \leqq +\infty$, the space of measurable functions f from I into X such that $\|f\|_X \in L^p(I)$. This is a Banach space for the norm

$$\|f\|_{L^p(I;X)} = \big| \|f\|_X \big|_{L^p(I)}.$$

Part I. Systems with a Polynomial Growth Nonlinearity

1. The equations and the semigroup. Let Ω denote an open bounded set of $\mathbb{R}^n$ with boundary Γ. We consider the following initial-boundary value problem involving a

vector function (u, v) from $\Omega\times\mathbb{R}^+$ into $\mathbb{R}^2$; (u, v) satisfies

$$\begin{aligned}&\frac{\partial u}{\partial t}-d\Delta u+h(x,u)+f(x,u,v)=0 \quad \text{in } \Omega\times\mathbb{R}^+,\\ &\frac{\partial v}{\partial t}+\sigma(x)v+g(x,u)=0 \quad \text{in } \Omega\times\mathbb{R}^+,\end{aligned} \tag{1.1}$$

together with the initial conditions

$$u(x,0)=u_0(x), \qquad v(x,0)=v_0(x), \tag{1.2}$$

and a boundary condition of either Dirichlet type

$$u=0 \quad \text{on } \Gamma, \tag{1.3}_1$$

or of Neumann type

$$\frac{\partial u}{\partial \nu}=0 \quad \text{on } \Gamma, \tag{1.3}_2$$

or of periodicity type

$$\Omega=(0,L)^n \quad u \text{ is } \Omega\text{-periodic.} \tag{1.3}_3$$

Here, the diffusion coefficient d is positive. Also, the functions h, f, σ and g are assumed to be twice continuously differentiable in all variables and to satisfy

$$\delta_1|u|^p-\delta_3\leqq h(x,u)u\leqq\delta_2|u|^p+\delta_3, \qquad p>2, \tag{1.4}$$

$$|f(x,u,v)|\leqq\delta_4(1+|u|^{p_1}+|v|), \qquad 0<p_1<p-1, \tag{1.5}$$

$$\sigma(x)\geqq\delta>0, \tag{1.6}$$

$$|g'_u(x,u)|\leqq\delta_5, \quad |g'_{x_i}(x,u)|\leqq\delta_5(1+|u|), \quad i=1,\cdots,n, \tag{1.7}$$

where the δ_i's are positive constants. We finally require the monotonicity assumption

$$(h'_u(x,u)+f'_u(x,u,v))\xi_1^2+f'_v(x,u,v)\xi_1\xi_2\geqq-\delta_6(\xi_1^2+\xi_2^2) \quad \forall(\xi_1,\xi_2)\in\mathbb{R}^2, \tag{1.8}$$

with $\delta_6\geqq 0$.

For the mathematical setting of problem (1.1)-$(1.3)_\alpha$ we introduce the functional spaces

$$H=\mathbb{L}^2(\Omega)=L^2(\Omega)^2, \qquad V=V_1\times L^2(\Omega),$$

where

$$V_1=\begin{cases}H_0^1(\Omega) & \text{for } \alpha=1,\\ H^1(\Omega) & \text{for } \alpha=2,\\ H_p^1(\Omega) & \text{for } \alpha=3,\end{cases}$$

where $H_p^k(\Omega)$, $k\in\mathbb{N}$, $\Omega=(0,L)^n$ denotes the space of functions that are locally in $H^k(\mathbb{R}^n)$ and are periodic with period L in each direction.[1]

With the above assumptions and using classical methods (see, for instance, Lions [16]), we can prove the following existence and uniqueness result.

[1] $u(x+Le_i)=u(x)$, where $(e_1,\cdots,e_n)$ denotes the canonical basis of $\mathbb{R}^n$.

PROPOSITION 1.1. *For (u_0, v_0) given in H, there exists a unique solution (u, v) of (1.1)-(1.3)$_\alpha$ satisfying*

$$(u, v) \in \mathscr{C}(\mathbb{R}^+, H),$$

$$u \in L^2(0, T; V_1) \cap L^p(Q_T), \quad Q_T = \Omega \times]0, T[\quad \text{for all } T > 0.$$

The mapping $(u_0, v_0) \to (u(t), v(t))$ is continuous on H.

2. The universal attractor. This section is devoted to the existence and to some properties of the universal attractor associated with (1.1)-(1.3)$_\alpha$. We first give in § 2.1, general definitions and results. We then prove in § 2.2 the existence of a universal attractor. Finally we conclude the section by deriving a regularity property of the universal attractor.

2.1. A general result on the existence of attractors. Let H be a subset of a Banach space $(E, |\ |)$ and let $S(t)$, $t \geqq 0$, be a semigroup of operators from H into itself. We recall that the *ω-limit set* of a subset $\mathscr{C}$ of H is defined by

$$\omega(\mathscr{C}) = \bigcap_{s \geqq 0} \overline{\bigcup_{t \geqq s} S(t)\mathscr{C}}^H.$$

A *functional invariant set* for the semigroup $S(t)$ is a set $\mathscr{X} \subset H$ such that

$$S(t)\mathscr{X} = \mathscr{X} \quad \forall t > 0.$$

If an invariant set $\mathscr{A}$, compact in H, possesses an open neighborhood $\mathscr{U}$ such that the image by $S(t)$ of any bounded subset of $\mathscr{U}$ converges to $\mathscr{A}$ as $t \to +\infty$, then $\mathscr{A}$ is called an *attractor* and the largest open set that contains $\mathscr{A}$ and that enjoys the same property as $\mathscr{U}$ is called the *basin of attraction* of $\mathscr{A}$. The *universal attractor*, if it exists, is the only attractor that admits H for basin of attraction, i.e., the image by $S(t)$ of any bounded set in H converges to $\mathscr{A}$ when $t \to +\infty$.

A set $\mathscr{B} \subset H$ is said to be *absorbing* for the semigroup $S(t)$ if, for every bounded set $\mathscr{B}_0$ of H, there exists $T = T(\mathscr{B}_0)$ such that $S(t)\mathscr{B}_0 \subset \mathscr{B}$, for all $t \geqq T(\mathscr{B}_0)$.

The following assumptions are made on the semigroup $S(t)$. First:

(2.1) $S(t)$ is continuous from H into itself, for all $t > 0$.

We also require that, for every t, $S(t) = S_1(t) + S_2(t)$, where the operators S_1, S_2 map H into E and satisfy the following:

(2.2) The operators $S_1(t)$ are uniformly compact in the following sense: for every bounded set $\mathscr{B}$ in H, there exists $t_0 \geqq 0$, such that $\bigcup_{t \geqq t_0} S_1(t)\mathscr{B}$ is relatively compact in E.

(2.3) For every bounded set $\mathscr{B} \subset H$, $r_{\mathscr{B}}(t) = \sup_{\varphi \in \mathscr{B}} |S_2(t)\varphi| \to 0$ as $t \to +\infty$.

We now recall a general result ensuring the existence of a universal attractor.

THEOREM 2.1. *We assume that (2.1)-(2.3) are satisfied and that there exists a bounded absorbing set $\mathscr{B}$. Then, the ω-limit set of $\mathscr{B}$, $\mathscr{A} = \omega(\mathscr{B})$, is the universal attractor for $S(t)$ in H. Furthermore, if H is convex, then $\mathscr{A}$ is connected.*

We conclude the section by a remark that will be useful in the sequel. Under the assumptions of Theorem 2.1, the universal attractor $\mathscr{A}$ is also the ω-limit set of $\mathscr{B}$ for the family of operators S_1, i.e.,

$$\mathscr{A} = \bigcap_{s \geqq 0} \overline{\bigcup_{t \geqq s} S_1(t)\mathscr{B}}^E.$$

Hence, $u \in \mathscr{A}$ if and only if there exist a sequence $u_l \in \mathscr{B}$ and a sequence $t_l \to +\infty$ such that

$$S_1(t_l)u_l \to u \quad \text{as } l \to +\infty. \tag{2.4}$$

The reader is referred to Temam [26] for more details and in particular for the proof of Theorem 2.1.

2.2. Existence of a universal attractor. We apply here the concepts and the results of § 2.1 to the semigroup associated with (1.1)-(1.3)$_\alpha$.

THEOREM 2.2. *Under assumptions* (1.4)-(1.8), *the semigroup* $\{S(t)\}_{t\in\mathbb{R}_+}$ *associated with* (1.1)-(1.3)$_\alpha$ *possesses a universal attractor* $\mathscr{A}$ *that is connected in H.*

Proof of Theorem 2.2.

(a) *Energy estimates and the existence of an absorbing set.* We multiply the first equation (1.1) by u and the second one by v and integrate over Ω. Using the Green formula and (1.3)$_\alpha$, we obtain

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}(|u|^2+|v|^2)+d\|u\|^2+\int_\Omega \sigma(x)v^2\,dx+\int_\Omega h(x,u)u\,dx \\ &\quad+\int_\Omega (f(x,u,v)u+g(x,u)v)\,dx=0. \end{aligned} \tag{2.5}$$

Due to (1.7), there exists a constant $\delta_7>0$ such that

$$|g(x,\xi)| \leqq \delta_7(1+|\xi|) \quad \forall \xi\in\mathbb{R}, \quad \forall x\in\Omega. \tag{2.6}$$

Using also (1.4)-(1.6), we deduce from (2.5) that

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}(|u|^2+|v|^2)+d\|u\|^2+\delta|v|^2+\delta_1\int_\Omega |u|^p\,dx \\ &\quad\leqq \delta_3|\Omega|+c_1\int_\Omega (|u|+|u|^{p_1+1})\,dx+c_1\int_\Omega |v|(1+|u|)\,dx, \end{aligned} \tag{2.7}$$

with $c_1=\delta_4+\delta_7$. The last integral in (2.7) is then majorized as follows:

$$c_1\int_\Omega |v|(1+|u|)\,dx \leqq \frac{\delta}{2}\int_\Omega |v|^2\,dx+\frac{c_1^2}{2\delta}\int_\Omega (|u|+1)^2\,dx.$$

Also, when we set $q=\max(p_1+1,2)$, there exists a constant $c_2>0$ such that

$$c_1\left(|\xi|+|\xi|^{p_1+1}+\frac{c_1^2}{2\delta}(1+|\xi|)^2\right) \leqq c_2(|\xi|^q+1) \quad \forall \xi\in\mathbb{R}.$$

Hence

$$\begin{aligned} c_1\int_\Omega \left(|u|+|u|^{p_1+1}+\frac{c_1^2}{2\delta}(|u|+1)^2\right)dx &\leqq c_2\int_\Omega |u|^q\,dx+c_2|\Omega| \\ &\leqq \frac{\delta_1}{4}\int_\Omega |u|^p\,dx+c_3 \quad \text{(using Young's inequality).} \end{aligned}$$

Combining the above inequalities, we infer from (2.7) that

$$\frac{1}{2}\frac{d}{dt}(|u|^2+|v|^2)+d\|u\|^2+\frac{\delta}{2}|v|^2+\frac{3\delta_1}{4}\int_\Omega |u|^p\,dx \leqq \delta_3|\Omega|+c_3. \tag{2.8}$$

Again using Young's inequality, we have

$$\frac{\delta}{2}\int_\Omega |u|^2\,dx \leqq \frac{\delta_1}{4}\int_\Omega |u|^p\,dx + c_4.$$

Hence (2.8) implies

$$\frac{d}{dt}(|u|^2+|v|^2)+2d\|u\|^2+\delta(|u|^2+|v|^2)+\delta_1\int_\Omega |u|^p\,dx \leqq c_5, \tag{2.9}$$

where $c_5 = 2(\delta_3|\Omega| + c_3 + c_4)$; (2.9) gives in particular

$$\frac{d}{dt}(|u|^2+|v|^2)+\delta(|u|^2+|v|^2) \leqq c_5,$$

which yields, using Gronwall's lemma,

$$|u(t)|^2+|v(t)|^2 \leqq (|u_0|^2+|v_0|^2)\exp(-\delta t)+\frac{c_5}{\delta}(1-\exp(-\delta t)). \tag{2.10}$$

We deduce from (2.10) that any ball of H centered at zero and of radius $\rho_2 > \rho_1 = (c_5/\delta)^{1/2}$ is an absorbing set. Indeed, if $\mathcal{B}$ is a bounded set of H, included in a ball $B(0, R)$ of H centered at zero and of radius R, then $S(t)\mathcal{B} \subset B(0, \rho_2)$ for $t \geqq T_0 = T_0(\mathcal{B}, \rho_2)$

$$T_0 = \frac{1}{\delta}\log\frac{R^2}{\rho_2^2-\rho_1^2}.$$

Let $\rho_2 > \rho_1$ and $r > 0$ be fixed. We infer from (2.9) after integrating in t that

$$2d\int_t^{t+r}\|u\|^2\,ds+\delta_1\int_t^{t+r}\int_\Omega |u|^p\,dx\,ds \leqq rc_5+|u(t)|^2+|v(t)|^2 \quad \forall t \geqq 0. \tag{2.11}$$

Hence, for $(u_0, v_0) \in \mathcal{B} \subset B(0, R)$ and $t \geqq T_0$,

$$2d\int_t^{t+r}\|u\|^2\,ds+\delta_1\int_t^{t+r}\int_\Omega |u|^p\,dx\,ds \leqq rc_5+\rho_2^2. \tag{2.12}$$

Again integrating (2.9) in t, we also get for $(u_0, v_0) \in \mathcal{B} \subset B(0, R)$,

$$\int_0^t \|u\|^2\,ds \leqq \frac{1}{2d}(c_5 t + R^2) \quad \forall t \geqq 0. \tag{2.13}$$

(b) The solution v of (1.1) can be written $v(t) = v_1(t) + v_2(t)$ with

$$\begin{aligned} v_1(x, t) &= -\int_0^t g(x, u(x, s))\, e^{-\sigma(x)(t-s)}\,ds, \\ v_2(x, t) &= v_0(x)\exp(-\sigma(x)t), \end{aligned} \tag{2.14}$$

and we define two families S_1, S_2 of operators from H into H by setting

$$\begin{aligned} S_1(t) &: (u_0, v_0) \to (u(t), v_1(t)), \\ S_2(t) &: (u_0, v_0) \to (0, v_2(t)). \end{aligned} \tag{2.15}$$

It is straightforward that $S_2(t)$ satisfies (2.3); indeed, for every bounded set $\mathcal{B} \subset H$,

$$r_{\mathcal{B}}(t) \leqq \exp(-\delta t)\sup_{\varphi\in\mathcal{B}}|\varphi|.$$

Our aim now is to check the uniform compactness of the operators $S_1(t)$ by using uniform in time a priori estimates on $u(t)$ and $v_1(t)$.

We multiply the first equation in (1.1) by $-\Delta u$ and integrate over Ω. Thanks to $(1.3)_\alpha$ and the Green formula, this gives the energy-type relation

$$\frac{1}{2}\frac{d}{dt}\|u\|^2+d|\Delta u|^2=\int_\Omega h(x,u)\,\Delta u\,dx+\int_\Omega f(x,u,v)\,\Delta u\,dx.$$

Due to (1.4), there exists a constant $\delta_8>0$ such that

$$|h(x,\xi)|\leqq\delta_8(1+|\xi|^{p-1})\quad\forall\xi\in\mathbb{R},\quad\forall x\in\Omega.$$

Hence, using also (1.5), we get, setting $c_6=\max(\delta_8,\delta_4)$,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u\|^2+d|\Delta u|^2&\leqq c_6\int_\Omega(2+|u|^{p_1}+|u|^{p-1}+|v|)|\Delta u|\,dx\\&\leqq\frac{d}{2}|\Delta u|^2+\frac{c_6^2}{2d}\int_\Omega(2+|u|^{p_1}+|u|^{p-1}+|v|)^2\,dx.\end{aligned}$$

Thus

$$\begin{aligned}\frac{d}{dt}\|u\|^2&\leqq\frac{4c_6^2}{d}\int_\Omega(4+|u|^{2p_1}+|u|^{2p-2}+|v|^2)\,dx\\&\leqq c_7\int_\Omega(1+|u|^{2p-2}+|v|^2)\,dx\quad(\text{since } p_1<p-1).\end{aligned}\tag{2.16}$$

□

We then apply the uniform Gronwall lemma, which we recall.

LEMMA 2.3. *Let g, h, y be three locally integrable scalar functions on $]t_0,\infty[$ satisfying*

$$\frac{dy}{dt}\in L^1_{\mathrm{loc}}(]t_0,+\infty[)\quad\text{and}\quad\frac{dy}{dt}\leqq gy+h\quad\text{for } t\geqq t_0,$$

$$\int_t^{t+r}g(s)\,ds\leqq a_1,\quad\int_t^{t+r}h(s)\,ds\leqq a_2,\quad\int_t^{t+r}y(s)\,ds\leqq a_3\quad\text{for } t\geqq t_0,$$

where r, a_1, a_2, a_3 are positive constants. Then

$$y(t+r)\leqq\left(\frac{a_3}{r}+a_2\right)\exp(a_1)\quad\forall t\geqq t_0.\tag{2.17}$$

The proof of this lemma can be found for instance in Foias, Manley, and Temam [9]. Thanks to the techniques of Lemma 2.5 below for $k=1$ and (2.12), we can check the existence of a constant c_8 such that for $(u_0,v_0)\in\mathcal{B}$ and $t\geqq T_0+r$

$$\int_t^{t+r}\int_\Omega|u|^{2p-2}\,dx\,ds\leqq c_8.$$

Hence, again using (2.12) and the existence of the absorbing set $B(0,\rho_2)$, we conclude that we can apply the uniform Gronwall lemma to (2.16) and we infer from (2.17) that

$$\begin{aligned}&\|u(t)\|^2\leqq c_9\quad\forall t\geqq T_0+2r,\\&c_9=\frac{1}{2dr}(rc_5+\rho_2^2)+c_7(r|\Omega|+c_8+r\rho_2^2).\end{aligned}\tag{2.18}$$

We now derive a time-uniform estimate of $v_1(t)$ in $H^1(\Omega)$. First, it is easy to deduce from (2.14) and (2.10) that there exists a constant $c_{10}=c_{10}(R)$ such that, for $(u_0, v_0)\in\mathcal{B}\subset B(0, R)$,

$$|v_1(t)|^2\leqq c_{10}(R)\quad \forall t\geqq 0. \tag{2.19}$$

We then set $w_j=\partial v_1/\partial x_j$, $j=1,\cdots,n$; w_j satisfies

$$\begin{aligned}&\frac{\partial w_j}{\partial t}+\sigma(x)w_j=-\sigma'_{x_j}(x)v_1-g'_{x_j}(x,u)-g'_u(x,u)\frac{\partial u}{\partial x_j},\\ &w_j(0)=0.\end{aligned} \tag{2.20}$$

The right-hand side of (2.20) satisfies the pointwise inequality (see (1.7))

$$\left|\sigma'_{x_j}(x)v_1+g'_{x_j}(x,u)+g'_u(x,u)\frac{\partial u}{\partial x_j}\right|\leqq(c_{11}+\delta_5)(1+|u|+|v_1|)+\delta_5\left|\frac{\partial u}{\partial x_j}\right|,$$

where $c_{11}=\max_{1\leqq j\leqq n}\max_{x\in\bar{\Omega}}|\sigma'_{x_j}(x)|$. Hence, by multiplying (2.20) by w_j and integrating over Ω, we find

$$\frac{1}{2}\frac{d}{dt}|w_j|^2+\delta|w_j|^2\leqq\frac{\delta}{2}|w_j|^2+\frac{1}{2\delta}\int_\Omega\left\{(c_{11}+\delta_5)(1+|u|+|v_1|)+\delta_5\left|\frac{\partial u}{\partial x_j}\right|\right\}^2 dx,$$

$$\frac{d}{dt}|w_j|^2+\delta|w_j|^2\leqq\frac{2}{\delta}(c_{11}+\delta_5)^2\int_\Omega(1+|u|+|v_1|)^2\,dx+\frac{2\delta_5^2}{\delta}\int_\Omega\left|\frac{\partial u}{\partial x_j}\right|^2 dx,$$

which, thanks to (2.19) and (2.10), for $(u_0, v_0)\in\mathcal{B}\subset B(0, R)$, gives

$$\begin{aligned}&\frac{d}{dt}|w_j|^2+\delta|w_j|^2\leqq c_{12}+\frac{2\delta_5^2}{\delta}\int_\Omega\left|\frac{\partial u}{\partial x_j}\right|^2 dx,\\ &c_{12}=c_{12}(R)=\frac{8}{\delta}(c_{11}+\delta_5)^2\left(|\Omega|+R^2+\frac{c_5}{\delta}+c_{10}\right).\end{aligned} \tag{2.21}$$

Summing (2.21) from $j=1$ to $j=n$, we finally obtain, for $(u_0, v_0)\in\mathcal{B}\subset B(0, R)$,

$$\frac{d}{dt}\|v_1\|^2+\delta\|v_1\|^2\leqq nc_{12}+\frac{2\delta_5^2}{\delta}\|u\|^2. \tag{2.22}$$

We then integrate this inequality; this gives, since $v_1(0)=0$,

$$\begin{aligned}\|v_1(t)\|^2&\leqq\frac{nc_{12}}{\delta}+\frac{2\delta_5^2}{\delta}\int_0^t\|u(s)\|^2\exp(\delta(s-t))\,ds\quad\forall t\geqq 0\\ &\leqq\frac{nc_{12}}{\delta}+\frac{2\delta_5^2}{\delta}\int_0^{T_0+2r}\|u(s)\|^2\exp(\delta(s-t))\,ds\\ &\quad+\frac{2\delta_5^2}{\delta}\int_{T_0+2r}^t\|u(s)\|^2\exp(\delta(s-t))\,ds\\ &\leqq\frac{nc_{12}}{\delta}+\frac{\delta_5^2}{\delta d}\{c_5(T_0+2r)+R^2\}+\frac{2\delta_5^2c_9}{\delta^2}\quad\forall t\geqq 0\end{aligned} \tag{2.23}$$

(using (2.13) and (2.18)).

The estimates (2.23), (2.19), (2.18), and (2.10) provide the uniform compactness of the operators S_1. Indeed, if $(u_0, v_0) \in \mathcal{B} \subset B(0, R)$ and $t \geqq T_0 + 2r$, then $S_1(t) \cdot (u_0, v_0)$ belongs to a set bounded in $\mathbb{H}^1(\Omega)$ independently of t and relatively compact in H.

In conclusion, the assumptions (2.1)-(2.3) are satisfied and we have proved the existence of a bounded absorbing set. Hence, Theorem 2.1 applies and gives Theorem 2.2. □

2.3. A regularity property of the attractor. The goal of this section is to prove the following regularity result.

THEOREM 2.4. *Under assumptions* (1.4)-(1.8), *the universal attractor* $\mathcal{A}$ *defined in Theorem* 2.2 *is bounded in* $\mathbb{L}^\infty(\Omega)$.

We start by proving the following weaker result.

LEMMA 2.5. *The attractor* $\mathcal{A}$ *defined in Theorem* 2.2 *is bounded in* $\mathbb{L}^q(\Omega)$, *for all* $q \in [1, +\infty[$.

Proof. Let $\alpha(k) = k(p-2)+2$. Since $\alpha(k) \to +\infty$, as $k \to +\infty$, it suffices to prove that $\mathcal{A}$ is bounded in $\mathbb{L}^{\alpha(k)}(\Omega)$, for all $k \in \mathbb{N}$. Let $r > 0$ be fixed; we shall prove by induction on k that

$$(2.24)_k \qquad \mathcal{A} \text{ is bounded in } \mathbb{L}^{\alpha(k)}(\Omega),$$

$$(2.25)_k \qquad \sup_{(u_0, v_0) \in \mathcal{A}} \int_t^{t+r} \int_\Omega |u|^{\alpha(k+1)}\, dx\, ds \leqq K \quad \forall t \geqq 0,$$

where (u, v) is the solution of (1.1)-$(1.3)_\alpha$; hereafter, we denote by K any constant that depends on the data and on k.

(a) $k = 0$; $\alpha(0) = 2$, and $\alpha(1) = p$; $(2.24)_0$ has been proved in Theorem 2.2. Let $(u_0, v_0) \in \mathcal{A}$; we infer from (2.11) that

$$\delta_1 \int_t^{t+r} \int_\Omega |u|^p\, dx\, ds \leqq rc_5 + |u(t)|^2 + |v(t)|^2 \quad \forall t \geqq 0$$

$$\leqq K \quad \text{(thanks to } (2.24)_0\text{)};$$

hence we have $(2.25)_0$.

(b) We now assume that $(2.24)_{k-1}$ and $(2.25)_{k-1}$ hold for $k \geqq 1$. In particular, there exists a constant $K > 0$ such that

$$(2.26) \qquad \int_\Omega |\tilde{v}|^{\alpha(k-1)}\, dx \leqq K \quad \forall (\tilde{u}, \tilde{v}) \in \mathcal{A}.$$

Let $(u_0, v_0) \in \mathcal{A}$. By multiplying the first equation in (1.1) by $|u|^{\alpha(k)-2}u$ and integrating over Ω, we obtain

$$(2.27) \qquad \begin{aligned} &\frac{1}{\alpha(k)} \frac{d}{dt} \int_\Omega |u|^{\alpha(k)}\, dx - d \int_\Omega \Delta u |u|^{\alpha(k)-2} u\, dx \\ &\quad + \int_\Omega h(x, u) |u|^{\alpha(k)-2} u\, dx + \int_\Omega f(x, u, v) |u|^{\alpha(k)-2} u\, dx = 0. \end{aligned}$$

Thanks to the Green formula and $(1.3)_\alpha$, we have

$$-\int_\Omega \Delta u |u|^{\alpha(k)-2} u\, dx = (\alpha(k)-1) \int_\Omega |\nabla u|^2 |u|^{\alpha(k)-2}\, dx \geqq 0.$$

Hence, using also (1.4) and (1.5), we deduce from (2.27)

$$\frac{1}{\alpha(k)}\frac{d}{dt}\int_\Omega |u|^{\alpha(k)}\,dx+\delta_1\int_\Omega |u|^{\alpha(k+1)}\,dx$$

$$\leqq \int_\Omega (\delta_3|u|^{\alpha(k)-2}+\delta_4|u|^{\alpha(k)-1}+\delta_4|u|^{p_1+\alpha(k)-1})\,dx$$

$$+\delta_4\int_\Omega |v|\,|u|^{\alpha(k)-1}\,dx$$

(since $p_1+\alpha(k)-1<\alpha(k+1)$),

$$\leqq \frac{\delta_1}{4}\int_\Omega |u|^{\alpha(k+1)}\,dx+K+\delta_4\int_\Omega |v|\,|u|^{\alpha(k)-1}\,dx, \tag{2.28}$$

$$\frac{1}{\alpha(k)}\frac{d}{dt}\int_\Omega |u|^{\alpha(k)}\,dx+\delta_1\frac{3}{4}\int_\Omega |u|^{\alpha(k+1)}\,dx$$

$$\leqq \delta_4\int_\Omega |v|\,|u|^{\alpha(k)-1}\,dx+K.$$

Let $\bar{p}=\alpha(k+1)/(\alpha(k)-1)$ and let $\bar{q}$ be such that $1/\bar{p}+1/\bar{q}=1$. With the Young inequality, the last integral in (2.28) is majorized as follows:

$$\delta_4\int_\Omega |v|\,|u|^{\alpha(k)-1}\,dx\leqq\frac{\delta_1}{4}\int_\Omega |u|^{\alpha(k+1)}\,dx+c_k'\int_\Omega |v|^{\bar{q}}\,dx.$$

Since $\bar{q}\leqq\alpha(k-1)$, using the Hölder inequality, we also have

$$\int_\Omega |v|^{\bar{q}}\,dx\leqq|\Omega|^{1-(\bar{q}/\alpha(k-1))}\left(\int_\Omega |v|^{\alpha(k-1)}\right)^{\bar{q}/\alpha(k-1)}$$

$$\leqq K \quad \text{(by (2.26))}.$$

Combining the above inequalities, we finally obtain

$$\frac{1}{\alpha(k)}\frac{d}{dt}\int_\Omega |u|^{\alpha(k)}\,dx+\frac{\delta_1}{2}\int_\Omega |u|^{\alpha(k+1)}\,dx\leqq K. \tag{2.29}$$

Thanks to the induction assumption $(2.25)_{k-1}$, we can apply the uniform Gronwall lemma to (2.29) and we conclude from (2.17) that there exists a constant K such that

$$\int_\Omega |u|^{\alpha(k)}\,dx\leqq K\quad\forall t\geqq r.$$

Since $S(r)\mathscr{A}=\mathscr{A}$, this implies

$$\int_\Omega |\tilde{u}|^{\alpha(k)}\,dx\leqq K\quad\forall(\tilde{u},\tilde{v})\in\mathscr{A}. \tag{2.30}$$

Then integrating (2.29) between t and $t+r$ and using (2.30), we find

$$\sup_{(u_0,v_0)\in\mathscr{A}}\int_t^{t+r}\int_\Omega |u|^{\alpha(k+1)}\,dx\,ds\leqq K\quad\forall t\geqq 0,$$

i.e., $(2.25)_k$.

It remains to check that

$$\int_\Omega |\tilde{v}|^{\alpha(k)}\,dx \leqq K \quad \forall (\tilde{u},\tilde{v}) \in \mathscr{A}. \tag{2.31}$$

Let $(\tilde{u},\tilde{v}) \in \mathscr{A}$. We claim that there exists a sequence $(u_{0l}, v_{0l}) \in \mathscr{A}$ and a sequence $t_l \to +\infty$ such that

$$S_1(t_l)\cdot(u_{0l}, v_{0l}) \to (\tilde{u},\tilde{v}) \quad \text{in } H \quad \text{as } l \to +\infty,$$

where $S_1(t)$ is given by (2.15). Indeed this follows from the functional invariance of $\mathscr{A}$ and the property (2.3) for $S_2(t) = S(t) - S_1(t)$.

Then, by (2.14),

$$\begin{aligned} |v_{1l}(t)|_{L^{\alpha(k)}(\Omega)} &\leqq \int_0^t |g(x,u_l)|_{L^{\alpha(k)}(\Omega)}\, e^{-\delta(t-s)}\,ds \quad (\delta \text{ given by (1.6)}) \\ &\leqq \delta_7 \int_0^t |1+|u_l||_{L^{\alpha(k)}(\Omega)}\, e^{-\delta(t-s)}\,ds \quad \text{(with (2.6))} \\ &\leqq \delta_7 K \frac{1}{\delta} \quad \forall t \geqq 0 \quad \text{(thanks to (2.30))}, \end{aligned} \tag{2.32}$$

which implies (2.31), since there exists a subsequence l_m such that

$$|\tilde{v}|_{L^{\alpha(k)}} \leqq \liminf_{m\to+\infty} |v_{1l_m}(t_{l_m})|_{L^{\alpha(k)}}.$$

The proof of Lemma 2.5 is therefore complete. □

Conclusion of the proof of Theorem 2.4. We now show that the attractor $\mathscr{A}$ is bounded in $\mathbb{L}^\infty(\Omega)$. Let $(\tilde{u},\tilde{v}) \in \mathscr{A}$ and let $\tilde{t} > 0$. Since $\mathscr{A}$ is a functional invariant set, there exists a solution (u, v) of (1.1)-(1.3)$_\alpha$ satisfying $(u_0, v_0) \in \mathscr{A}$ and $(u(\tilde{t}), v(\tilde{t})) = (\tilde{u},\tilde{v})$. Introducing the semigroup $\Sigma(t)$ associated with the linear operator $\partial/\partial t - d\Delta + I$ and with the boundary condition (1.3)$_\alpha$, it is classical that u can be written as

$$u(t) = \Sigma(t)u_0 + \int_0^t \Sigma(t-s)\{-h(x,u(s)) - f(x,u(s),v(s)) + u(s)\}\,ds \quad \text{for } t \geqq 0. \tag{2.33}$$

The semigroup $\Sigma(t)$ satisfies the regularity property (see Rothe [22])

$$|\Sigma(t)\varphi|_{L^\infty(\Omega)} \leqq cm(t)^{-1/2}\, e^{-\lambda t}|\varphi|_{L^n(\Omega)},$$

where $m(t) = \min(1, t)$, λ is the smallest eigenvalue of the operator $-d\Delta + I$ associated with the boundary condition (1.3)$_\alpha$, and c is a positive constant. Also, by Lemma 2.5, there exists a constant $K > 0$ such that

$$|u|_{L^n(\Omega)},\ |-h(x,u) - f(x,u,v) + u|_{L^n(\Omega)} \leqq K.$$

Hence, we deduce from (2.33) that

$$\begin{aligned} |u(t)|_{L^\infty(\Omega)} &\leqq cK\left\{ e^{-\lambda t}m(t)^{-1/2} + \int_0^t e^{-\lambda(t-s)}m(t-s)^{-1/2}\,ds \right\} \quad \forall t \geqq 0 \\ &\leqq cK\left\{ m\left(\frac{\tilde{t}}{2}\right)^{-1/2} + 2 + \frac{1}{\lambda} \right\} \quad \forall t \geqq \frac{\tilde{t}}{2}. \end{aligned}$$

In particular, $\tilde{u} = u(\tilde{t})$ satisfies

$$|\tilde{u}|_{L^\infty(\Omega)} \leqq cK\left\{m\left(\frac{\tilde{t}}{2}\right)^{-1/2} + 2 + \frac{1}{\lambda}\right\} \quad \forall (\tilde{u}, \tilde{v}) \in \mathscr{A}.$$

Finally, the bound on $|\tilde{v}|_{L^\infty(\Omega)}$ follows from the one on $|\tilde{u}|_{L^\infty(\Omega)}$ as in Lemma 2.5 above. This concludes the proof of Theorem 2.4. □

3. Dimension of the universal attractor. Our aim is now to prove the finite dimensionality of the attractor introduced before. We start by giving a few results borrowed from Constantin, Foias, and Temam [6] and Ghidaglia and Temam [10] (§ 3.1). We then derive in § 3.2 estimates of the Hausdorff and fractal dimensions of the attractor. Finally, we conclude the section by applying our results to an example (§ 3.3).

3.1. Some general results on the dimension of functional invariant sets. Let E be a Hilbert space (norm $|\cdot|$) and let $\mathscr{X}$ be a compact functional invariant set for a semigroup $\{S(t)\}_{t \geqq 0}$. We assume the following for all $t \geqq 0$:

(3.1) $S(t)$ is uniformly differentiable in $\mathscr{X}$, which means that for every $u \in \mathscr{X}$ there exists a linear operator $L = L(t, u) \in \mathscr{L}(E)$ such that

$$\sup_{\substack{u, v \in \mathscr{X} \\ 0 < |v-u| \leqq \varepsilon}} \frac{|S(t)v - S(t)u - L(t, u) \cdot (v-u)|}{|v-u|} \to 0 \quad \text{as } \varepsilon \to 0,$$

(3.2)
$$\sup_{u \in \mathscr{X}} |L(t, u)|_{\mathscr{L}(E)} < +\infty.$$

For $L \in \mathscr{L}(E)$ and $N \in \mathbb{N}^*$, we denote by $\omega_N(L)$ the norm of the exterior product $\Lambda^N L$ in $\Lambda^N E$:

$$\omega_N(L) = \sup_{\substack{\xi^1, \cdots, \xi^N \in E \\ |\xi^j| \leqq 1}} |L\xi^1 \wedge \cdots \wedge L\xi^N|.$$

We then introduce for $N \in \mathbb{N}^*$

$$\bar{\omega}_N(t) = \sup_{u_0 \in \mathscr{X}} \omega_N(L(t, u_0)), \qquad t \geqq 0,$$

$$\Pi_N = \lim_{t \to +\infty} \bar{\omega}_N(t)^{1/t}$$

and we define the uniform Lyapunov numbers μ_N by

$$\mu_1 = \log \Pi_1, \quad \mu_N = \log \Pi_N - \log \Pi_{N-1}, \quad N \geqq 2.$$

A general result on the Hausdorff and fractal dimensions of functional invariant sets states the following.

THEOREM 3.1. *Under assumptions* (3.1), (3.2), *if, for some integer* $N \geqq 1$, $\mu_1 + \cdots + \mu_N < 0$, *then the Hausdorff dimension of* $\mathscr{X}$ *is less than or equal to* N *and the fractal dimension of* $\mathscr{X}$ *is less than or equal to*

$$N \cdot \max_{1 \leqq l \leqq N-1} \left(1 + \frac{(\mu_1 + \cdots + \mu_l)_+}{|\mu_1 + \cdots + \mu_N|}\right).$$

The reader is referred to Constantin, Foias and Temam [6] for the proof of this result when the linear operators $L(t, u)$ are assumed to be compact and to Ghidaglia and Temam [10] for the extension to the noncompact case that we use here. We conclude this section by recalling the definitions of the Hausdorff and fractal dimensions

(Federer [7], Mandelbrot [17]). The d-dimensional Hausdorff measure of $\mathscr{X}$ is the number

$$\mu_H(\mathscr{X}, d) = \lim_{\varepsilon \to 0} \mu_H(\mathscr{X}, d, \varepsilon)$$

where

$$\mu_H(\mathscr{X}, d, \varepsilon) = \inf \sum_i r_1^d,$$

the infimum being taken for all the coverings of $\mathscr{X}$ by balls of radii $r_1 \leqq \varepsilon$. There exists $d = d_H(\mathscr{X}) \in [0, +\infty]$ such that $\mu_H(\mathscr{X}, d) = 0$ for $d > d_H(\mathscr{X})$ and is equal to ∞ for $d < d_H(\mathscr{X})$; $d_H(\mathscr{X})$ is the Hausdorff dimension of $\mathscr{X}$.

The fractal dimension of $\mathscr{X}$ is

$$d_F(\mathscr{X}) = \inf \{d > 0, \mu_F(\mathscr{X}, d) = 0\}$$

where

$$\mu_F(\mathscr{X}, d) = \limsup_{\varepsilon \to 0} \varepsilon^d n_{\mathscr{X}}(\varepsilon),$$

and $n_{\mathscr{X}}(\varepsilon)$ is the minimum number of balls of radii ε necessary to cover $\mathscr{X}$. Since $\mu_F(\mathscr{X}, d) \geqq \mu_H(\mathscr{X}, d)$, it is clear that the fractal dimension of a set is larger than or equal to its Hausdorff dimension, the converse being false in general.

3.2. Estimate of the dimension of the attractor. We now return to the universal attractor $\mathscr{A}$ defined in Theorem 2.2. According to Theorem 2.4, there exists a constant $\alpha > 0$ such that

$$|\tilde{u}|_{L^\infty(\Omega)}, |\tilde{v}|_{L^\infty(\Omega)} \leqq \alpha \quad \forall (\tilde{u}, \tilde{v}) \in \mathscr{A}$$

and we introduce

$$c_3 = \inf_{\substack{x \in \bar{\Omega} \\ |u| \leqq \alpha, |v| \leqq \alpha}} (h'_u(x, u) + f'_u(x, u, v)), \tag{3.3}$$

$$c_4 = \sup_{\substack{x \in \bar{\Omega} \\ |u| \leqq \alpha, |v| \leqq \alpha}} |f'_v(x, u, v) + g'_u(x, u)|. \tag{3.4}$$

(Note that $c_3 \geqq -\delta_6$ given by (1.8).) The aim of this section is to prove Theorem 3.2.

THEOREM 3.2. *Under assumptions* (1.4)-(1.8), *the Hausdorff and fractal dimensions of the universal attractor* $\mathscr{A}$ *defined in Theorem* 2.2 *are finite. Moreover, they are both bounded by*

$$c\left\{1 + \frac{d}{\delta |\Omega|^{2/n}} + \frac{|\Omega|}{d^{n/2}}\left(\delta^{n/2} + \frac{1}{\delta}(c_3^-)^{n_1} + \frac{1}{\delta^{n_1+1}} c_4^{2n_1}\right)\right\} \tag{3.5}$$

where c *denotes a constant depending on* n *and the shape of* Ω;[2] $n_1 = 1 + n/2$, $c_3^- = \max(0, -c_3)$, *and* δ, c_3, c_4 *are given by* (1.6), (3.3), *and* (3.4), *respectively.*

Remark 3.3. In the course of the proof of Theorem 3.2 below, we derive the following sharper bound of the Hausdorff and fractal dimensions:

$$c\left\{1 + \frac{d}{\delta |\Omega|^{2/n}} + \frac{\delta^{n/2}|\Omega|}{d^{n/2}} + \frac{c_5}{\delta\, d^{n/2}}\right\},$$

[2] This means that the constant is invariant by homothety and translation of Ω.

where

$$c_5 = \liminf_{t\to+\infty} \sup_{(u_0,v_0)\in\mathscr{A}} \left\{ \frac{1}{t}\int_0^t \int_\Omega \left\{ ((h'_u(x,u)+f'_u(x,u,v))^-)^{n_1} + \frac{|f'_v(x,u,v)+g'_u(x,u)|^{2n_1}}{\delta^{n_1}} \right\} dx\, ds \right\}, \tag{3.6}$$

and c is as in Theorem 3.2.

Theorem 3.2 is proved by using the general result recalled in § 3.1. We start by checking the assumptions (3.1) and (3.2) for the semigroup $S(t)$ associated with (1.1)-$(1.3)_\alpha$.

LEMMA 3.4. *For every $t_0>0$, $s(t_0)$ is uniformly differentiable on $\mathscr{A}$. Its differential at (u_0, v_0) is the linear operator on $H: (\xi,\eta)\to L(t_0,u_0,v_0)$. $(\xi,\eta)=(U(t_0), V(t_0))$, where $(U(t_0), V(t_0))$ is the value at time t_0 of the solution $(U(t), V(t))$ of the linearized problem*

$$\frac{\partial U}{\partial t} - d\Delta U + [h'_u(x,u)+f'_u(x,u,v)]U + f'_v(x,u,v)V = 0, \tag{3.7}$$

$$\frac{\partial V}{\partial t} + \sigma(x)V + g'_u(x,u)U = 0,$$

$$U(0)=\xi, \qquad V(0)=\eta, \tag{3.8}$$

$$U \text{ satisfies } (1.3)_\alpha,$$

where (u, v) is the solution of (1.1)-$(1.3)_\alpha$. Furthermore, we have

$$\sup_{(u_0,v_0)\in\mathscr{A}} |L(t_0,u_0,v_0)|_{\mathscr{L}(H)} < +\infty.$$

Proof of Lemma 3.4. The proof of this technical result is left to the reader. We only note here that it relies essentially on Theorem 2.4. □

For $(u_0, v_0)\in\mathscr{A}$, let us denote by $\mathfrak{A}(u, v)$ the linear operator from $H^2(\Omega)\times L^2(\Omega)$ into H occurring in (3.7), (3.8), i.e.,

$$\mathfrak{A}(u,v)\cdot(U,V) = (-d\Delta U + (h'_u+f'_u)U + f'_v V,\ \sigma V + g'_u U).$$

For $N\geqq 1$, we introduce

$$q_N = \limsup_{t\to+\infty} \left(\inf_{(u_0,v_0)\in\mathscr{A}} \frac{1}{t}\int_0^t \inf_{\text{rank } Q=N} \operatorname{Tr}(\mathfrak{A}(u,v)\circ Q)\, ds \right), \tag{3.9}$$

where Q denotes any orthogonal projector in H of rank N such that $QH\subset W$ with

$$W = \{(U,V)\in H^2(\Omega)\times L^2(\Omega),\ U \text{ satisfies } (1.3)_\alpha\}.$$

Returning to the definitions of § 3.1, it can be shown exactly as in [6] that

$$\mu_1+\cdots+\mu_N \leqq -q_N, \qquad N\geqq 1.$$

Thus, we infer from Theorem 3.1 that, if $q_N>0$ for some N, then the Hausdorff dimension of $\mathscr{A}$ is less than or equal to N and its fractal dimension is majorized by

$$N\cdot \max_{1\leqq l\leqq N-1} \left(1+\frac{(-q_l)_+}{q_N}\right). \tag{3.10}$$

We next estimate the q_N's. Let Q be an orthogonal projector in H of rank N such that $QH \subset W$ and let $\{(\varphi^j, \psi^j)\}_{j\in\mathbb{N}}$ be an orthonormal basis of H such that $(\varphi^1, \psi^1), \cdots, (\varphi^N, \psi^N)$ is a basis of QH. Then

$$
\begin{aligned}
\operatorname{Tr} \mathfrak{A}(u, v) \circ Q &= \sum_{j=1}^{N} (\mathfrak{A}(u, v) \cdot (\varphi^j, \psi^j), (\varphi^j, \psi^j)) \\
&= \sum_{j=1}^{N} \Big(d\|\varphi^j\|^2 + \int_\Omega \sigma(x)\psi^{j^2}\, dx \\
&\qquad + \int_\Omega (h'_u(x, u) + f'_u(x, u, v))\varphi^{j^2}\, dx \\
&\qquad + \int_\Omega (f'_v(x, u, v) + g'_u(x, u))\varphi^j\psi^j\, dx \Big) \\
&\geqq d \sum_{j=1}^{N} \|\varphi^j\|^2 + \delta \sum_{j=1}^{N} |\psi^j|^2 - \int_\Omega (h'_u + f'_u)^- \rho(x)\, dx \\
&\qquad + \sum_{j=1}^{N} \int_\Omega (f'_v + g'_u)\varphi^j\psi^j\, dx \quad \text{(by (1.6))},
\end{aligned}
\tag{3.11}
$$

where we have set

$$
\rho(x) = \sum_{j=1}^{N} (\varphi^j(x))^2.
$$

After majorizing the last integral as follows:

$$
\left| \sum_{j=1}^{N} \int_\Omega (f'_v + g'_u)\varphi^j\psi^j\, dx \right| \leqq \frac{\delta}{2} \sum_{j=1}^{N} \int_\Omega |\psi^j|^2\, dx + \frac{1}{2\delta} \int_\Omega (f'_v + g'_u)^2 \rho(x)\, dx,
$$

we have that (3.11) gives

$$
\operatorname{Tr} \mathfrak{A}(u, v) \circ Q \geqq d \sum_{j=1}^{N} \|\varphi^j\|^2 + \frac{\delta}{2} \sum_{j=1}^{N} |\psi^j|^2 - \int_\Omega \alpha_1(x, t)\rho(x)\, dx \tag{3.12}
$$

where

$$
\alpha_1(x, t) = (h'_u + f'_u)^- + \frac{1}{2\delta}(f'_v + g'_u)^2. \tag{3.13}
$$

We next apply the generalized Lieb-Thirring inequalities (see Ghidaglia, Marion, and Temam [11] to the suborthonormal[3] family $(\varphi^j)_{1\leqq j\leqq N}$: there exist two constants K_1 and K_2 depending on n and the shape of Ω such that

$$
\sum_{j=1}^{N} \|\varphi^j\|^2 \geqq K_1 \int_\Omega \rho(x)^{1+(2/n)}\, dx - \frac{K_2}{L^2} \int_\Omega \rho(x)\, dx, \tag{3.14}
$$

where L denotes the diameter of Ω. Thus, we infer from (3.12) that

$$
\begin{aligned}
\operatorname{Tr} \mathfrak{A}(u, v) \circ Q &\geqq \frac{\delta}{2} \sum_{j=1}^{N} (|\varphi^j|^2 + |\psi^j|^2) \\
&\qquad + dK_1 \int_\Omega \rho(x)^{1+(2/n)}\, dx - \int_\Omega \left(\frac{\delta}{2} + \frac{dK_2}{L^2} + \alpha_1(x, t) \right) \rho(x)\, dx.
\end{aligned}
\tag{3.15}
$$

[3] We say (cf. [11]) that the family $\varphi^j \in L^2(\Omega)$, $1 \leqq j \leqq N$, is suborthonormal if $\sum_{i,j=1}^{N} \xi_i\xi_j \int_\Omega \varphi^i\varphi^j\, dx \leqq \sum_{k=1}^{N} \xi_k^2$, for all $\xi \in \mathbb{R}^N$.

Since the family (φ^j, ψ^j) is orthonormal, we then have

$$\sum_{j=1}^{N} (|\varphi^j|^2 + |\psi^j|^2) = N.$$

Also, with the Young inequality, the last integral in (3.15) can be majorized as follows:

$$\int_\Omega \left(\frac{\delta}{2} + \frac{dK_2}{L^2} + \alpha_1(x,t)\right)\rho(x)\,dx \leqq \frac{dK_1}{2}\int_\Omega \rho(x)^{1+(2/n)}\,dx + \alpha_2(t),$$

(3.16)

$$\alpha_2(t) = \frac{2}{n+2}\left(\frac{4n}{(n+2)\,dK_1}\right)^{n/2}\int_\Omega \left(\frac{\delta}{2} + \frac{dK_2}{L^2} + \alpha_1(x,t)\right)^{n_1} dx, \qquad n_1 = 1 + \frac{n}{2}.$$

Combining these inequalities, we finally obtain

$$\operatorname{Tr}\mathfrak{A}(u,v)\circ Q \geqq \frac{\delta}{2}N + \frac{dK_1}{2}\int_\Omega \rho(x)^{1+(2/n)}\,dx - \alpha_2(t)$$

(3.17)

$$\geqq \frac{\delta}{2}N - \alpha_2(t).$$

Hence, we conclude from (3.17) and definition (3.9) that

$$q_N \geqq \frac{\delta}{2}N - \alpha, \tag{3.18}$$

where

$$\alpha = \liminf_{t\to+\infty} \sup_{(u_0,v_0)\in\mathscr{A}} \frac{1}{t}\int_0^t \alpha_2(t)\,dt. \tag{3.19}$$

Let now $\bar{N}$ be the integer such that

$$\frac{\delta}{2}(\bar{N}-1) \leqq 2\alpha < \frac{\delta}{2}\bar{N}. \tag{3.20}$$

Then $q_{\bar{N}} > 0$. As observed before, the Hausdorff dimension is majorized by $\bar{N}$ and, by (3.10) and (3.18),

$$d_F(\mathscr{A}) \leqq \bar{N} \max_{1\leqq l\leqq \bar{N}-1}\left(1 + \frac{\alpha}{\alpha}\right) = 2\bar{N}.$$

To conclude, we express the bound on $\bar{N}$ given by (3.20). We find

$$\bar{N} \leqq 1 + \frac{4\alpha}{\delta}$$

$$\leqq c\left(1 + \frac{d}{\delta|\Omega|^{2/n}} + \frac{\delta^{n/2}|\Omega|}{d^{n/2}} + \frac{c_5}{\delta\, d^{n/2}}\right) \quad \text{(by (3.19), (3.16), and (3.13))},$$

where c_5 is given by (3.6), and c depends on n and the shape of Ω. This gives (3.5) by using the constants c_3 and c_4 given by (3.3) and (3.4). The proof of Theorem 3.2 is therefore complete. □

3.3. An example. Here we apply our results to the system

$$\frac{\partial u}{\partial t} - d\Delta u + h(u) + \sigma v = 0, \qquad \frac{\partial v}{\partial t} + \delta v + \gamma u = 0, \tag{3.21}$$

where $\delta > 0$, $\delta, \gamma \in \mathbb{R}$, and h is a polynomial of odd degree greater than 1 with a positive leading coefficient.

Assumptions (1.5)-(1.8) are satisfied. Theorems 2.2 and 3.2 apply and give the existence of a universal attractor that describes the long-time behavior of the solutions of (3.21) (the boundary condition is either the Dirichlet or the Neumann or the periodic boundary condition). This attractor has finite Hausdorff and fractal dimensions and these dimensions moreover are bounded by

$$c\left[1+\frac{d}{\delta|\Omega|^{2/n}}+\frac{|\Omega|}{d^{n/2}}\left\{\delta^{n/2}+\frac{1}{\delta}(c_3^-)^{n_1}+\frac{1}{\delta^{n_1+1}}|\sigma+\gamma|^{2n_1}\right\}\right],$$

where c denotes a constant depending on n and the shape of Ω, $n_1=1+n/2$, and

$$c_3^-=\max(0,-c_3), \qquad c_3=\min_{u\in\mathbb{R}} h'(u).$$

Part II. Systems with an Invariant Region.

In the second part of this work we consider partly dissipative systems of reaction-diffusion equations admitting a positively invariant region. The precise assumptions on the equations are stated in § 4. Then, in § 5, we prove the existence of a universal attractor and we derive an estimate of the dimension of the attractor. Finally, in § 5, we apply our results to several classical equations borrowed from mathematical biology, physics and chemistry.

4. The equations and the semigroup. We denote again by Ω an open bounded set of $\mathbb{R}^n$ with boundary Γ. Let $m=m_1+m_2$ and let D be a positive diagonal matrix of diffusion coefficients

$$D=\operatorname{diag}(d_1,\cdots,d_{m_1}), \qquad d_i>0.$$

The system of ordinary differential equations coupled with partial differential equations to be considered involves a vector function (u, v) from $\Omega\times\mathbb{R}_+$ into $\mathbb{R}^{m_1}\times\mathbb{R}^{m_2}$ and takes the form

$$\begin{aligned}&\frac{\partial u}{\partial t}-D\Delta u+f(x,u,v)=0,\\ &\frac{\partial v}{\partial t}+G(x,u)v+g(x,u)=0,\end{aligned} \tag{4.1}$$

where $f=(f_1,\cdots,f_{m_1})$ is a function of class $\mathscr{C}^2$ on $\bar{\Omega}\times\mathbb{R}^m$, $g=(g_1,\cdots,g_{m_2})$ is a function of class $\mathscr{C}^2$ on $\bar{\Omega}\times\mathbb{R}^{m_1}$, and $G=(g_{kl})$ is a square matrix of order m_2 whose coefficients are functions of class $\mathscr{C}^2$ on $\bar{\Omega}\times\mathbb{R}^{m_1}$. We supplement (4.1) with the initial conditions

$$u(x,0)=u_0(x), \qquad v(x,0)=v_0(x) \quad \text{in } \Omega, \tag{4.2}$$

and a boundary condition of either Dirichlet type

$$u=0 \quad \text{on } \Gamma, \tag{4.3}_1$$

of Neumann type

$$\frac{\partial u}{\partial \nu}=0 \quad \text{on } \Gamma, \tag{4.3}_2$$

or of periodicity type

$$\Omega=]0,L[^n, \quad u \text{ is } \Omega\text{-periodic.} \tag{4.3}_3$$

The following assumptions are made on the initial boundary value problem (4.1)-(4.3)$_\alpha$. First:

(4.4) There exists a closed convex region $\mathscr{D} \subset \mathbb{R}^m$ that is positively invariant. Moreover, $\mathscr{D}$ is compact for $\alpha = 2, 3$.

This means that any solution (u, v), with boundary data and initial data $(u_0(x), v_0(x))$ in $\mathscr{D}$ for every (or almost every) x in Ω, satisfies $(u(x, t), v(x, t)) \in \mathscr{D}$ for all $t > 0$ for which the solution exists. We refer to Chueh, Conley, and Smoller [3] for the derivation of sufficient and of necessary and sufficient conditions that guarantee the existence of a positively invariant region (see also Tartar [25], Sermange [23]).

For the mathematical setting of the problem, we introduce the functional spaces

$$H = L^2(\Omega, \mathscr{D}) = \{(u, v) \in \mathbb{L}^2(\Omega), (u(x), v(x)) \in \mathscr{D} \text{ for a.e. } x \in \Omega\},$$
$$V = V_1 \times (L^2(\Omega))^{m_2},$$

where

$$V_1 = \begin{cases} H_0^1(\Omega)^{m_1} & \text{if } \alpha = 1, \\ H^1(\Omega)^{m_1} & \text{if } \alpha = 2, \\ H_p^1(\Omega)^{m_1} & \text{if } \alpha = 3. \end{cases}$$

We assume that (4.1)-(4.3)$_\alpha$ is well posed, for (u_0, v_0) in H, in the following sense:

(4.5) For $(u_0, v_0) \in H$, (4.1)-(4.3)$_\alpha$ possesses a unique solution (u, v) for all time, $(u(t), v(t)) \in H$, for all t, $(u, v) \in L^2(0, T; V)$, for all $T > 0$. The mapping $(u_0, v_0) \to (u(t), v(t))$ is continuous in H. Moreover, if $(u_0, v_0) \in V$, then $u \in L^2(0, T; H^2(\Omega)^{m_1})$, for all $T > 0$.

Finally, we assume the following:

(4.6) There exists $\delta > 0$ such that, for all $(x, u, v) \in \bar{\Omega} \times \mathscr{D}$,

$$\sum_{k,l=1}^{m_2} g_{kl}(x, u)\xi_k \xi_l \geqq \delta \sum_{k=1}^{m_2} \xi_k^2 \quad \forall \xi \in \mathbb{R}^{m_2}.$$

(4.7) f, g and the partial derivatives of order one and two of f, g and G are bounded on $\bar{\Omega} \times \mathscr{D}$.

We also set

$$c_1 = \sup_{(x,u,v) \in \bar{\Omega} \times \mathscr{D}} |f(x, u, v)|, \tag{4.8}$$

$$c_2 = \sup_{(x,u,v) \in \bar{\Omega} \times \mathscr{D}} |g(x, u)|. \tag{4.9}$$

5. The universal attractor.

5.1. Existence of a universal attractor. Our goal in this section is to prove Theorem 5.1.

THEOREM 5.1. *Under assumptions* (4.4)-(4.7), *the semigroup* $S(t)$ *associated with* (4.1)-(4.3)$_\alpha$ *possesses a universal attractor* $\mathscr{A}$ *that is connected in* H.

Proof of Theorem 5.1. The proof relies on technical a priori estimates and on the general results given in § 2.1.

We first note that when the positively invariant region $\mathscr{D}$ is compact (Neumann and periodic boundary conditions), the existence of an absorbing set is straightforward and Theorem 5.1 then follows by using the same arguments as in step (b) below for the Dirichlet case. So, in the sequel, we restrict ourselves to the Dirichlet boundary condition.

(a) *Existence of an absorbing set.* We multiply the first equation (4.1) by u and integrate over Ω. Using the Green formula and $(4.3)_\alpha$, we obtain

$$\frac{1}{2}\frac{d}{dt}|u|^2+\sum_{i=1}^{m_1} d_i\|u_i\|^2+\int_\Omega f(x,u,v)\cdot u\,dx=0.$$

Introducing the constant

$$d_0=\min_{1\leqq i\leqq m_1} d_i, \tag{5.1}$$

and using the constant c_1 defined in (4.8), we get

$$\frac{1}{2}\frac{d}{dt}|u|^2+d_0\|u\|^2\leqq c_1|\Omega|^{1/2}|u|. \tag{5.2}$$

Thanks to the Poincaré inequality, there exists a constant $c_3>0$ such that $|u|\leqq c_3\|u\|$ for all $u\in V_1$. Hence, we infer from (5.2) that

$$\frac{1}{2}\frac{d}{dt}|u|^2+\frac{d_0}{c_3^2}|u|^2\leqq c_1|\Omega|^{1/2}|u|\leqq\frac{1}{2}\frac{d_0}{c_3^2}|u|^2+\frac{c_1^2c_3^2|\Omega|}{2d_0},$$

$$\frac{d}{dt}|u|^2+\frac{d_0}{c_3^2}|u|^2\leqq\frac{c_1^2c_3^2|\Omega|}{d_0},$$

which gives, by integration,

$$|u(t)|^2\leqq|u_0|^2\exp\left(-\frac{d_0}{c_3^2}t\right)+\frac{c_1^2c_3^4|\Omega|}{d_0^2}\left(1-\exp\left(-\frac{d_0}{c_3^2}t\right)\right)\quad\forall t\geqq 0. \tag{5.3}$$

Let $\rho_2>\rho_1=c_1c_3^2|\Omega|^{1/2}/d_0$ be fixed and let $\mathcal{B}$ be any bounded set of H included in a ball $B(0,R)$ of H centered at zero and of radius R. We deduce from (5.3) that, if $(u_0,v_0)\in\mathcal{B}$, then

$$|u(t)|\leqq\rho_2\quad\forall t\geqq T_1(\mathcal{B},\rho_2), \tag{5.4}$$

where

$$T_1=\frac{c_3^2}{d_0}\log\frac{R^2}{\rho_2^2-\rho_1^2}.$$

Also, by integrating (5.2) between t and $t+r$, $r>0$ fixed, we have

$$\int_t^{t+r}\|u\|^2\,ds\leqq\frac{c_1|\Omega|^{1/2}}{d_0}\int_t^{t+r}|u|\,ds+\frac{|u(t)|^2}{2d_0},$$

and, if $(u_0,v_0)\in\mathcal{B}\subset B(0,R)$ and $t\geqq T_1(\mathcal{B},\rho_2)$,

$$\int_t^{t+r}\|u\|^2\,ds\leqq c_4,\qquad c_4=\frac{1}{d_0}rc_1|\Omega|^{1/2}\rho_2+\frac{\rho_2^2}{2d_0}. \tag{5.5}$$

Integrating again (5.2) between zero and t and using (5.3), we get furthermore that, if $(u_0,v_0)\in\mathcal{B}\subset B(0,R)$,

$$\int_0^t\|u\|^2\,ds\leqq\frac{1}{d_0}c_1|\Omega|^{1/2}\sqrt{R^2+\rho_1^2}\,t+\frac{R^2}{2d_0}\quad\forall t\geqq 0. \tag{5.6}$$

Now, multiplying the second equation in (4.1) by v and integrating over Ω, we obtain

$$\frac{1}{2}\frac{d}{dt}|v|^2+\int_\Omega G(x,u)v\cdot v\,dx+\int_\Omega g(x,u)\cdot v\,dx=0.$$

Hence, using the constant δ defined in (4.6), we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}|v|^2+\delta|v|^2&\leqq-\int_\Omega g(x,u)\cdot v\,dx\\&\leqq c_2|\Omega|^{1/2}|v|\quad(c_2\text{ given by (4.9)})\\&\leqq\frac{\delta}{2}|v|^2+\frac{c_2^2|\Omega|}{2\delta},\\\frac{d}{dt}|v|^2+\delta|v|^2&\leqq\frac{c_2^2|\Omega|}{\delta},\end{aligned}$$

which gives

$$|v(t)|^2\leqq|v_0|^2\exp(-\delta t)+\frac{c_2^2|\Omega|}{\delta^2}(1-\exp(-\delta t))\quad\forall t\geqq0. \tag{5.7}$$

Let $\rho_4>\rho_3=c_2|\Omega|^{1/2}/\delta$ be fixed. It follows from (5.7) that if $(u_0,v_0)\in\mathcal{B}\subset B(0,R)$ then

$$|v(t)|\leqq\rho_4\quad\forall t\geqq T_2(\mathcal{B}_0,\rho_4), \tag{5.8}$$

where

$$T_2(\mathcal{B}_0,\rho_4)=\frac{1}{\delta}\log\frac{R^2}{\rho_4^2-\rho_3^2}.$$

We conclude immediately from (5.4) and (5.8) that the ball of H of radius $\{\rho_2^2+\rho_4^2\}^{1/2}$ and of center zero is absorbing in H.

(b) We now write $v(t)=v_1(t)+v_2(t)$, where $v_1(t)$ is the unique solution of

$$\frac{\partial v_1}{\partial t}+G(x,u)v_1+g(x,u)=0, \tag{5.9}$$

$$v_1(0)=0, \tag{5.10}$$

and $v_2(t)$ is the unique solution of

$$\frac{\partial v_2}{\partial t}+G(x,u)v_2=0, \tag{5.11}$$

$$v_2(0)=v_0. \tag{5.12}$$

We define two families S_1, S_2 of nonlinear operators from H into $E=\mathbb{L}^2(\Omega)$ by setting

$$S_1(t):(u_0,v_0)\to(u(t),v_1(t)),\qquad S_2(t):(u_0,v_0)\to(0,v_2(t)).$$

Our aim is to derive properties (2.2) and (2.3) for this decomposition. It is easy to check (2.3). Indeed, multiplying (5.11) by v_2 and integrating over Ω, we get

$$\frac{1}{2}\frac{d}{dt}|v_2|^2+\int_\Omega G(x,u)v_2\cdot v_2\,dx=0.$$

Hence, by (4.6),

$$\frac{1}{2}\frac{d}{dt}|v_2|^2+\delta|v_2|^2\leqq 0,$$

which yields

$$|v_2(t)|^2\leqq \exp(-2\delta t)|v_0|^2.$$

This implies immediately (2.3).

Next, we multiply (4.1) by $-\Delta u$ and integrate over Ω. We obtain

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2+(D\Delta u,\Delta u)=\int_\Omega f(x,u,v)\cdot\Delta u\,dx.$$

Thus, using the constants d_0 and c_1 defined in (5.1) and (4.8), we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u(t)\|^2+d_0|\Delta u|^2&\leqq c_1|\Omega|^{1/2}|\Delta u|\\ &\leqq \frac{d_0}{2}|\Delta u|^2+\frac{c_1^2|\Omega|}{2d_0},\end{aligned}\tag{5.13}$$

$$\frac{d}{dt}\|u(t)\|^2+d_0|\Delta u|^2\leqq\frac{c_1^2|\Omega|}{d_0}.$$

Thanks to (5.5), we can apply the uniform Gronwall lemma to (5.13). We conclude from (2.17) that if $(u_0,v_0)\in\mathcal{B}\subset B(0,R)$, then

$$\|u(t)\|^2\leqq\frac{c_4}{r}+\frac{rc_1^2|\Omega|}{d_0}\quad\text{for } t\geqq T_1+r.\tag{5.14}$$

Now we aim to establish time uniform estimates for $v_1(t)$. We first claim that

$$|v_1(t)|_{\mathbb{L}^\infty(\Omega)}\leqq\frac{2c_2}{\delta}\quad\forall t\geqq 0,\tag{5.15}$$

where δ and c_2 are given by (4.6) and (4.9). Indeed, for any given $q>2$, let us multiply (5.19) by $|v_1|^{q-2}v_1$ and integrate over Ω. We get

$$\frac{1}{q}\frac{d}{dt}\int_\Omega|v_1|^q\,dx+\int_\Omega|v_1|^{q-2}G(x,u)v_1\cdot v_1\,dx=-\int_\Omega|v_1|^{q-2}g(x,u)\cdot v_1\,dx.\tag{5.16}$$

Due to (4.6), we have

$$\int_\Omega|v_1|^{q-2}G(x,u)v_1\cdot v_1\,dx\geqq\delta\int_\Omega|v_1|^q\,dx.\tag{5.17}$$

Also, by (4.9),

$$\begin{aligned}\left|\int_\Omega|v_1|^{q-2}g(x,u)\cdot v_1\,dx\right|&\leqq c_2\int_\Omega|v_1|^{q-1}\,dx\\ &\leqq\frac{\delta}{2}\int_\Omega|v_1|^q\,dx+c_5(q)\quad\text{(by Young's inequality),}\end{aligned}\tag{5.18}$$

where

$$c_5(q)=\frac{1}{q}\,c_2^q|\Omega|\left(\frac{2(q-1)}{\delta q}\right)^{q-1}.$$

Hence, combining (5.16)-(5.18), we obtain

$$\frac{1}{q}\frac{d}{dt}\int_\Omega |v_1|^q\,dx+\frac{\delta}{2}\int_\Omega |v_1|^q\,dx\leqq c_5(q).$$

This implies by integrating, since $v_1(0)=0$,

$$|v_1(t)|_{\mathbb{L}^q(\Omega)}\leqq\left(\frac{2c_5(q)}{\delta}\right)^{1/q}\quad\forall t\geqq 0,$$

which yields (5.15) by taking the limit $q\to+\infty$.

We now derive an estimate of $\|v_1(t)\|$. For $j=1,\cdots,n$, we set $w_j=\partial v_1/\partial x_j$; w_j satisfies

$$\frac{\partial w_j}{\partial t}+G(x,u)w_j+\frac{\partial G}{\partial x_j}v_1+\sum_{i=1}^{m_1}\frac{\partial u_i}{\partial x_j}\frac{\partial G}{\partial u_i}v_1+\frac{\partial g}{\partial x_j}+\sum_{i=1}^{m_1}\frac{\partial u_i}{\partial x_j}\frac{\partial g}{\partial u_i}=0. \tag{5.19}$$

Thanks to the assumption (4.7) and to (5.15), it is easy to see that the vector functions

$$\frac{\partial G}{\partial x_j}v_1,\ \frac{\partial G}{\partial u_i}v_1,\ \frac{\partial g}{\partial x_j},\ \frac{\partial g}{\partial u_i},\quad j=1,\cdots,n,\quad i=1,\cdots,m_1,$$

are bounded in $\mathbb{L}^\infty((0,+\infty)\times\Omega)$. Hence, if we multiply (5.19) by w_j and integrate over Ω, we find

$$\frac{1}{2}\frac{d}{dt}|w_j|^2+\int_\Omega G(x,u)w_j\cdot w_j\,dx\leqq c_6\int_\Omega |w_j|\left(\left|\frac{\partial u}{\partial x_j}\right|+1\right)dx,$$

where c_6 denotes a constant depending only on the data. Using (4.6), we then have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}|w_j|^2+\delta|w_j|^2&\leqq c_6\int_\Omega |w_j|\left(\left|\frac{\partial u}{\partial x_j}\right|+1\right)dx\\&\leqq\frac{\delta}{2}|w_j|^2+\frac{c_6^2}{\delta}\left\{|\Omega|+\left|\frac{\partial u}{\partial x_j}\right|^2\right\},\\\frac{d}{dt}|w_j|^2+\delta|w_j|^2&\leqq\frac{2c_6^2}{\delta}\left\{|\Omega|+\left|\frac{\partial u}{\partial x_j}\right|^2\right\}.\end{aligned}$$

Summing these inequalities from $j=1$ up to $j=n$, we finally obtain

$$\frac{d}{dt}\|v_1\|^2+\delta\|v_1\|^2\leqq\frac{2nc_6^2|\Omega|}{\delta}+\frac{2c_6^2}{\delta}\|u\|^2. \tag{5.20}$$

This inequality is similar to (2.22) and, as in the proof of Theorem 2.2, we can infer from (5.6), (5.14), and (5.20) that there exists a constant $c_7=c_7(R)$ such that, if $(u_0,v_0)\in\mathcal{B}\subset B(0,R)$,

$$\|v_1(t)\|\leqq c_7\quad\forall t\geqq 0. \tag{5.21}$$

It is now easy to conclude the proof of Theorem 5.1. It follows from (5.21), (5.15), (5.14) and (5.3) that the operators S_1 are uniformly compact. Hence, the assumptions (2.1)-(2.3) are satisfied and there exists a bounded absorbing set in H. Theorem 2.1 applies and yields Theorem 5.1. □

Remark 5.2. The proof of Theorem 5.1 actually shows that

$$|v|_{\mathbb{L}^\infty(\Omega)} \leqq \frac{2c_2}{\delta} \quad \forall (u, v) \in \mathscr{A}. \tag{5.22}$$

This result, which follows from (5.15) and (2.4), will be useful in the next section.

5.2. Estimate of the dimension of the attractor. We introduce the following constants:

$$c_3 = \sup_{(x,u,v)\in\bar{\Omega}\times\mathscr{D}} \left\{ \sum_{i=1}^{m_1} \left| \frac{\partial f}{\partial u_i} \right|^2 \right\}^{1/2}, \tag{5.23}$$

$$c_4 = \sup_{(x,u,v)\in\bar{\Omega}\times\mathscr{D}} \left\{ \sum_{l=1}^{m_2} \left| \frac{\partial f}{\partial v_l} \right|^2 + \sum_{i=1}^{m_1} \left(\left| \frac{\partial g}{\partial u_i} \right|^2 + \frac{c_2^2}{\delta^2} \left| \frac{\partial G}{\partial u_i} \right|^2 \right) \right\}^{1/2}, \tag{5.24}$$

where δ and c_2 are given by (4.6) and (4.9). Our aim is to prove Theorem 5.3.

THEOREM 5.3. *Under assumptions* (4.4)-(4.7), *the universal attractor* $\mathscr{A}$ *defined in Theorem* 5.1 *has finite Hausdorff and fractal dimensions. Moreover, both dimensions are bounded by*

$$c\left\{ 1 + \frac{d_0}{\delta |\Omega|^{2/n}} + \frac{|\Omega|}{d_0^{n/2}} \left(\delta^{n/2} + \frac{1}{\delta} c_3^{n_1} + \frac{1}{\delta^{n_1+1}} c_4^{2n_1} \right) \right\}, \tag{5.25}$$

where c *denotes a constant depending on* n, m_1 *and the shape of* Ω; $n_1 = 1 + n/2$ *and* δ, d_0, c_3, c_4 *are, respectively, given by* (4.6), (5.1), (5.23), *and* (5.24).

Remarks 5.4. (1) A remark similar to Remark 3.3 could be made here; the details are left to the reader.

(2) In all the examples of § 6 below, the positively invariant region $\mathscr{D}$ is bounded in the v direction; i.e., there exists $\alpha > 0$ such that $|v| \leqq \alpha$, for all $(u, v) \in \mathscr{D}$. It follows that the universal attractor $\mathscr{A}$ satisfies

$$|v|_{\mathbb{L}^\infty(\Omega)} \leqq \alpha \quad \forall (u, v) \in \mathscr{A}.$$

Using this bound (instead of (5.22)) and the same arguments as in the proof of Theorem 5.3, we can derive another estimate of the Hausdorff and fractal dimensions of the attractor:

$$c\left\{ 1 + \frac{d_0}{\delta |\Omega|^{2/n}} + \frac{|\Omega|}{d_0^{n/2}} \left(\delta^{n/2} + \frac{1}{\delta} c_3^{n_1} + \frac{1}{\delta^{n_1+1}} c_4'^{2n_1} \right) \right\}, \tag{5.25}'$$

where

$$c_4' = \sup_{(x,u,v)\in\bar{\Omega}\times\mathscr{D}} \left\{ \sum_{l=1}^{m_2} \left| \frac{\partial f}{\partial v_l} \right|^2 + \sum_{i=1}^{m_1} \left(\left| \frac{\partial g}{\partial u_i} \right|^2 + \alpha^2 \left| \frac{\partial G}{\partial u_i} \right|^2 \right) \right\}^{1/2}.$$

We shall use the estimate (5.25)′ in all the examples of § 6.

Proof of Theorem 5.3. The proof follows the same steps as for Theorem 3.2 and most of the computations in the proof of Theorem 3.2 will be adapted here. For $(u_0, v_0) \in \mathscr{A}$, let us denote by $\mathfrak{A}(u, v)$ the linear operator from $(H^2(\Omega))^{m_1} \times (L^2(\Omega))^{m_2}$ into $E = \mathbb{L}^2(\Omega)$ defined by

$$\mathfrak{A}(u, v) \cdot (U, V) = (\mathfrak{A}_1(u, v) \cdot (U, V), \mathfrak{A}_2(u, v) \cdot (U, V)),$$

$$\mathfrak{A}_1(u, v) \cdot (U, V) = -D\Delta U + \sum_{i=1}^{m_1} \frac{\partial f}{\partial u_i}(x, u, v) U_i + \sum_{l=1}^{m_2} \frac{\partial f}{\partial v_l}(x, u, v) V_l,$$

$$\mathfrak{A}_2(u, v) \cdot (U, V) = G(x, u) V + \sum_{i=1}^{m_1} \frac{\partial G}{\partial u_i}(x, u) v U_i + \sum_{i=1}^{m_1} \frac{\partial g}{\partial u_i}(x, u) U_i,$$

where (u, v) is the solution of (4.1)-$(4.3)_\alpha$.

We first show that, for every $t_0 > 0$, $S(t_0)$ is uniformly differentiable on $\mathscr{A}$. Its differential at (u_0, v_0) is the linear operator on $E: (\xi, \eta) \to L(t_0, u_0, v_0) \cdot (\xi, \eta) = (U(t_0), V(t_0))$, where $(U(t_0), V(t_0))$ is the value at time t_0 of the solution $(U(t), V(t))$ of the following linearized problem:

$$\dot{U} + \mathfrak{A}_1(u, v) \cdot (U, V) = 0, \qquad \dot{V} + \mathfrak{A}_2(u, v) \cdot (U, V) = 0,$$
$$U(0) = \xi, \qquad V(0) = \eta,$$
$$U \text{ satisfies } (4.3)_\alpha$$

(we omit the details).

We then define

$$q_N = \limsup_{t \to +\infty} \left(\inf_{(u_0, v_0) \in \mathscr{A}} \frac{1}{t} \int_0^t \inf_{\text{rank } Q = N} \operatorname{Tr} (\mathfrak{A}(u, v) \circ Q)\, ds \right),$$

where Q denotes any orthogonal projector of rank N in E such that

$$QE \subset \{(U, V) \in (H^2(\Omega))^{m_1} \times (L^2(\Omega))^{m_2},\ U \text{ satisfies the boundary condition } (4.3)_\alpha\}.$$

As in Theorem 3.2, if $q_N > 0$ for some integer N, then the Hausdorff dimension of $\mathscr{A}$ is majorized by N and its fractal dimension is bounded by (3.10).

To estimate the q_N's, let Q be an orthogonal projector satisfying the above conditions and let $\{(\varphi^j, \psi^j)\}_{j \in \mathbb{N}}$ be an orthonormal basis of E such that $(\varphi^1, \psi^1), \cdots, (\varphi^N, \psi^N)$ is a basis of the image of Q. Then

$$\begin{aligned} \operatorname{Tr} \mathfrak{A}(u, v) \circ Q &= \sum_{j=1}^N (\mathfrak{A}(u, v)(\varphi^j, \psi^j), (\varphi^j, \psi^j)) \\ &= \sum_{j=1}^N \left(\sum_{i=1}^{m_1} d_i \|\varphi_i^j\|^2 + \int_\Omega G\psi^j \cdot \psi^j\, dx + \sum_{i,k=1}^{m_1} \int_\Omega \frac{\partial f_k}{\partial u_i} \varphi_i^j \varphi_k^j\, dx \right) \\ &\quad + \sum_{j=1}^N \sum_{i=1}^{m_1} \sum_{l=1}^{m_2} \int_\Omega \left(\frac{\partial f_i}{\partial v_l} + \left(\frac{\partial G}{\partial u_i} v \right)_l + \frac{\partial g_l}{\partial u_i} \right) \varphi_i^j \psi_l^j\, dx. \end{aligned} \tag{5.26}$$

Using the constants d_0 and δ defined in (5.1) and (4.6), we have

$$\sum_{j=1}^N \left(\sum_{i=1}^{m_1} d_i \|\varphi_i^j\|^2 + \int_\Omega G\psi^j \cdot \psi^j\, dx \right) \geqq d_0 \sum_{j=1}^N \|\varphi^j\|^2 + \delta \sum_{j=1}^N |\psi^j|^2.$$

Also

$$\left| \sum_{j=1}^N \sum_{i,k=1}^{m_1} \int_\Omega \frac{\partial f_k}{\partial u_i} \varphi_i^j \varphi_k^j\, dx \right| \leqq c_3 \int_\Omega \rho(x)\, dx \quad (c_3 \text{ given by (5.23)}),$$

where we have set

$$\rho(x) = \sum_{j=1}^N |\varphi^j(x)|^2.$$

Using (5.22) and the constant c_4 defined in (5.24), we have that the last integral in (5.26) is majorized as follows:

$$\begin{aligned} &\left| \sum_{j=1}^N \sum_{i=1}^{m_1} \sum_{l=1}^{m_2} \int_\Omega \left(\frac{\partial f_i}{\partial v_l} + \left(\frac{\partial G}{\partial u_i} v \right)_l + \frac{\partial g_l}{\partial u_i} \right) \varphi_i^j \psi_l^j\, dx \right| \\ &\quad \leqq c_5 c_4 \sum_{j=1}^N |\varphi^j|\, |\psi^j| \quad (c_5 \text{ is a universal constant}) \\ &\quad \leqq \frac{\delta}{2} \sum_{j=1}^N |\varphi^j|^2 + \frac{c_5^2 c_4^2}{2\delta} \int_\Omega \rho(x)\, dx. \end{aligned}$$

Combining the above inequalities, from (5.26) we infer that

$$\operatorname{Tr} \mathfrak{A}(u, v) \circ Q \geqq d_0 \sum_{j=1}^{N} \|\varphi^j\|^2 + \frac{\delta}{2} \sum_{j=1}^{N} \|\psi^j|^2 - \left(c_3 + \frac{c_5^2 c_4^2}{2\delta}\right) \int_\Omega \rho(x)\, dx.$$

This estimate is analogous to (3.12). Moreover, the generalized Lieb-Thirring inequalities for the vector functions $(\varphi^j)_{1 \leqq j \leqq N}$ read like the ones (see (3.14)) for the scalar functions with constants K_1, K_2 that depend on n, m_1, and the shape of Ω. Hence, we can conclude the proof of Theorem 5.3 thanks to computations similar to the ones from (3.12) in the proof of Theorem 3.2; the details are left to the reader. □

6. Examples. We describe some examples of reaction-diffusion systems satisfying the hypotheses (4.4)-(4.7).

Example 6.1. *Hodgkin-Huxley equations.* This system, proposed by Hodgkin and Huxley [13], describes the nerve impulse transmission. Here $n = 1$, $\Omega = (0, L)$ and the system is of the form (4.1) with $m_1 = 1$, $m_2 = 3$:

$$\begin{aligned} \frac{\partial u}{\partial t} &= d \frac{\partial^2 u}{\partial x^2} - f(u, v_1, v_2, v_3), \\ \frac{\partial v_1}{\partial t} &= k_1(u)(h_1(u) - v_1), \\ \frac{\partial v_2}{\partial t} &= k_2(u)(h_2(u) - v_2), \\ \frac{\partial v_3}{\partial t} &= k_3(u)(h_3(u) - v_3). \end{aligned} \tag{6.1}$$

We have

$$f(u, v) = -\gamma_1 v_1^2 v_2 (\sigma_1 - u) - \gamma_2 v_3^4 (\sigma_2 - u) - \gamma_3 (\sigma_3 - u),$$

with $\gamma_i > 0$, $\sigma_1 > \sigma_3 > 0 > \sigma_2$. Furthermore, $d > 0$, k_i, h_i are $\mathscr{C}^\infty$ functions and satisfy $k_i > 0$, $1 > h_i > 0$, $i = 1, 2, 3$; u represents the electrical potential in the nerve, while v_1, v_2, v_3 represent chemical concentrations and vary thus between zero and one.

It is proved in Chueh, Conley, and Smoller [3] and Sermange [23] that any rectangle

$$\mathscr{D} = \{(u, v_1, v_2, v_3),\ \alpha_0 \leqq u \leqq \alpha_1,\ 0 \leqq v_i \leqq 1,\ i = 1, 2, 3\}$$

is an invariant region provided $\alpha_1 \geqq \delta_1$ (> 0), $\alpha_0 \leqq \delta_2$ (< 0). Assumptions (4.6) and (4.7) are obviously satisfied and (4.5) follows from the existence of the compact invariant region $\mathscr{D}$ (see [18] for more details).

Theorems 5.1 and 5.3 apply and give the existence of a universal attractor $\mathscr{A}$ whose Hausdorff and fractal dimensions are finite (the boundary condition is the Dirichlet, the Neumann, or the periodic boundary condition). This attractor describes the long-time behavior of the impulse transmission in the nerve. Its Hausdorff and fractal dimensions are more precisely bounded by

$$c\left(1 + \frac{1}{\delta L^2} d + \frac{1}{d^{1/2}} L\left(\delta^{1/2} + \frac{1}{\delta} c_3^{3/2} + \frac{1}{\delta^{5/2}} c_4^3\right)\right),$$

where c is a universal constant and

$$\delta = \min_{1 \leqq i \leqq 3} \min_{\alpha_0 \leqq u \leqq \alpha_1} k_i(u),$$

$$c_3 = \max_{(u, v) \in \mathscr{D}} \left|\frac{\partial f}{\partial u}(u, v)\right|,$$

$$c_4 = \max_{(u,v)\in\mathscr{D}} \left(\sum_{i=1}^{3} \left(\left| \frac{\partial f}{\partial v_i}(u,v) \right|^2 + |k_i(u)h_i'(u)|^2 + |k_i'(u)|^2 \right) \right)^{1/2}.$$

Example 6.2. *FitzHugh-Nagumo equations.* These equations introduced by FitzHugh [8] and Nagumo, Arimoto, and Yosimzawa [21] are also intended to describe the signal transmission across axons; they read as follows:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + h(u) - v, \qquad \frac{\partial v}{\partial t} = \sigma u - \delta v.$$

Here $n = 1$, $\Omega = (0, L)$, $m_1 = m_2 = 1$, $\sigma > 0$, $\delta > 0$ and

$$h(u) = -u(u-\beta)(u-1), \qquad 0 < \beta < \tfrac{1}{2}.$$

As in (6.1), u represents the electrical potential in the axon, but v has a more complicated interpretation.

It is proved in Chueh, Conley, and Smoller [3] that any rectangle

$$\mathscr{D} = \{(u, v), -\alpha_1 \leqq u \leqq \alpha_2, -\alpha_3 \leqq v \leqq \alpha_4\}$$

is an invariant region provided the edges $v = \alpha_4$ and $v = -\alpha_3$ of $\mathscr{D}$ are, respectively, included in the half spaces $\sigma u - \delta v < 0$ and $\sigma u - \delta v > 0$ and the edges $u = \alpha_2$ and $u = -\alpha_1$ are on both sides of the zero set of $h(u) - v$. Assumptions (4.6) and (4.7) are obviously satisfied and (4.5) is proved as in Example 6.1 above.

Theorem 5.1 and 5.3 apply and give the existence of a universal attractor having finite Hausdorff and fractal dimensions (the boundary condition is either the Dirichlet or the Neumann or the periodic boundary condition). Moreover, both dimensions are bounded by

$$c\left(1 + \frac{1}{\delta L^2} + L\left(\delta^{1/2} + \frac{1}{\delta} c_3^{3/2} + \frac{1}{\delta^{5/2}} c_4^3\right)\right),$$

where c is a universal constant and

$$c_3 = \max_{-\alpha_1 \leqq u \leqq \alpha_2} |h'(u)|, \qquad c_4 = \{1 + \sigma^2\}^{1/2}.$$

Example 6.3. *A system of solid combustion type.* In the theory of combustion of solids, there arises the following system:

$$\frac{\partial u}{\partial t} = \Delta u + \sigma v h(u), \qquad \frac{\partial v}{\partial t} = -\sigma v h(u) + k(x).$$

Here $n \geqq 1$, $m_1 = m_2 = 1$; $\sigma > 0$, $k(x)$ is a $\mathscr{C}^2$ function on $\bar{\Omega}$ with $k \geqq 0$ and we have

$$h(u) = \exp\left(-\frac{\beta(1-u)}{1+\gamma(u-1)}\right),$$

where $\beta > 0$, $\gamma \in]0, 1[$; u represents the temperature while v represents the concentration of the solid reactant. The original model where $k = 0$ (see Matkowsky and Sivashinsky [20]) leads to a trivial attractor; here we introduce a source term of reactant $k \neq 0$. From a physical point of view, a useful variant of the model consists in assuming a constant flux of reactant across the boundary (Clavin et al. [4], Clavin [5]); the corresponding problem, which requires some modifications of the above framework, will be studied in another paper [19].

Using the classical truncation method, we can show that the rectangle

$$\mathscr{D} = \{(u, v), u \geqq 0, 0 \leqq v \leqq \alpha\}, \qquad \alpha = \frac{1}{\sigma} \exp\left(\frac{\beta}{1-\gamma}\right) \sup_{x\in\bar{\Omega}} |k(x)|$$

is a positively invariant region. Also, assumptions (4.6) and (4.7) are satisfied and (4.5) is proved thanks to an estimate of $|u(t)|_{L^\infty(\Omega)}$.

Theorems 5.1 and 5.3 apply and give the existence of a universal attractor with finite Hausdorff and fractal dimensions (for a Dirichlet boundary condition). These dimensions are majorized by

$$c\left(1+\frac{1}{\delta|\Omega|^{2/n}}+|\Omega|\left(\delta^{n/2}+\frac{1}{\delta}c_3^{n_1}+\frac{1}{\delta^{n_1+1}}c_4^{2n_1}\right)\right),$$

where c depends on n and the shape of Ω, $n_1=1+n/2$, and

$$\delta=\sigma\exp\left(-\frac{\beta}{1-\gamma}\right),\qquad c_3=\frac{\gamma^2}{\beta}\exp\left(\frac{\beta}{\gamma(1-\gamma)}\right)\sup_{x\in\bar{\Omega}}|k(x)|,$$

$$c_4=\left(\sigma^2\exp\left(\frac{2\beta}{\gamma}\right)+c_3^2\right)^{1/2}.$$

Example 6.4. *Feld-Noyes equations.* The equations of this last example serve as a model for the Belousov-Zhabotinskii reactions in chemical kinetics (cf. Howard and Kopell [14], Hastings and Murray [12]). Here, $n\geqq 1$, $m_1=1$, $m_2=2$ and (u, v_1, v_2) satisfy

$$\begin{aligned}\frac{\partial u}{\partial t}&=d\Delta u+\alpha(v_1-uv_1+u-\beta u^2),\\ \frac{\partial v_1}{\partial t}&=\frac{1}{\alpha}(\gamma v_2-v_1-uv_1),\\ \frac{\partial v_2}{\partial t}&=\sigma(u-v_2),\end{aligned}\tag{6.2}$$

where α, β, γ, σ are positive constants. The constant d can usually be greater than or equal to zero, but we impose $d>0$; u, v_1, v_2 represent chemical concentrations.

For convenience, we introduce the new unknowns $w_1=v_1$, $w_2=\xi v_2$, where $\xi>0$ will be chosen later. Then the system (6.2) becomes

$$\begin{aligned}\frac{\partial u}{\partial t}&=d\Delta u+\alpha(w_1-uw_1+u-\beta u^2),\\ \frac{\partial w_1}{\partial t}&=\frac{1}{\alpha}\left(\frac{\gamma}{\xi}w_2-w_1-uw_1\right),\\ \frac{\partial w_2}{\partial t}&=\sigma(\xi u-w_2).\end{aligned}$$

It is proved by Chueh, Conley, and Smoller [3] that

$$\mathcal{D}=\{(u, w_1, w_2), 0\leqq u\leqq a, 0\leqq w_1\leqq b, 0\leqq w_2\leqq \xi c\}$$

is invariant provided $a>\max(1, \beta^{-1})$, $c>a$, $b>\gamma c$. Assumption (4.7) is obvious and (4.5) is proved as in Example 6.1; (4.6) is satisfied provided $\gamma^2<4\sigma\alpha\xi^2$ and we choose $\xi=\gamma/(2\sigma\alpha)^{1/2}$.

Theorems 5.1 and 5.3 apply and give the existence of a universal attractor whose Hausdorff and fractal dimensions are bounded by

$$c\left(1+\frac{1}{\delta|\Omega|^{2/n}}+\frac{|\Omega|}{d^{n/2}}\left(\delta^{n/2}+\frac{1}{\delta}c_3^{n_1}+\frac{1}{\delta^{n_1+1}}c_4^{2n_1}\right)\right)$$

where c depends on n and the shape of Ω, $n_1 = 1 + n/2$, and

$$\delta = \min_{u \in [0,a]} \frac{1}{2}\left(\frac{1+u}{\alpha} + \sigma - \left(\left(\frac{1+u}{\alpha} - \sigma\right)^2 + \frac{2\sigma}{\alpha}\right)^{1/2}\right),$$

$$c_3 = \alpha(b + 1 + \beta a),$$

$$c_4 = \left(\alpha^2(1+a)^2 + \sigma^2 + \frac{1}{\alpha^2}\left(b^2 + \frac{\gamma^2 c^2}{\sigma\alpha}\right)\right)^{1/2}.$$

REFERENCES

[1] A. V. BABIN AND M. I. VISHIK, *Regular attractors of semigroups and evolution equations*, J. Math. Pures Appl., 62 (1983), pp. 441-491.

[2] ———, *Attractors of partial differential equations and estimates of their dimension*, Russian Math. Surveys, 38 (1983), pp. 151-213.

[3] K. N. CHUEH, C. C. CONLEY, AND J. A. SMOLLER, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373-392.

[4] P. CLAVIN, B. DENET, J. MONTEILLIER, AND P. PELCÉ, *Study of reaction-diffusion waves representative of reverse combustion in porous combustible media*, J. Méc. Théor. Appl., numéro spécial (1986), pp. 173-192.

[5] P. CLAVIN, personal communication.

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

[7] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[8] R. FITZHUGH, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445-466.

[9] C. FOIAS, O. MANLEY, AND R. TEMAM, *Attractors for the Bénard problem: existence and physical bounds on their fractal dimension*, Nonlinear Anal.: Theory Meth. Appl., 11 (1987), pp. 939-967.

[10a] J. M. GHIDAGLIA AND R. TEMAM, *Propriétés des attracteurs associés à des équations hyperboliques non linéaires amorties*, C.R. Acad. Sci. Paris Série I, 300 (1985), pp. 185-188.

[10b] ———, *Attractors for damped nonlinear hyperbolic equations*, J. Math. Pures Appl., 66 (1987), pp. 273-319.

[11] J. M. GHIDAGLIA, M. MARION, AND R. TEMAM, *Sur quelques inégalités fonctionnelles*, C.R. Acad. Sci. Paris Sér. I (1987), pp. 287-290, and *Generalization of the Sobolev-Lieb-Thirring inequalities and applications to the dimension of attractors*, Differential and Integral Equations, 1 (1988), pp. 1-21.

[12] S. HASTINGS AND J. MURRAY, *The existence of oscillatory solutions in the Field-Noyes model for the Belousov-Zhabotinskii reaction*, SIAM J. Appl. Math., 28 (1975), pp. 678-688.

[13] A. L. HODGKIN AND A. P. HUXLEY, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 500-544.

[14] L. HOWARD AND N. KOPELL, *Plane wave solutions to reaction-diffusion equations*, Stud. Appl. Math., 52 (1973), pp. 291-328.

[15] N. KOPELL AND D. RUELLE, *Bounds on complexity in reaction-diffusion systems*, SIAM J. Appl. Math., 46 (1986), pp. 68-80.

[16] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[17] B. MANDELBROT, *Fractals: Form, Chance and Dimension*, W. H. Freeman, San Francisco, 1977.

[18] M. MARION, *Attractors for reaction-diffusion equations: existence and estimate of their dimension*, Appl. Anal., 25 (1987), pp. 101-147.

[19] ———, manuscript in preparation.

[20] B. J. MATKOWSKY AND G. I. SIVASHINSKY, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Appl. Math., 35 (1978), pp. 465-478.

[21] J. NAGUMO, S. ARIMOTO, AND S. YOSIMZAWA, *An active pulse transmission line simulating nerve axon*, Proc. I.R.E., 50 (1964), pp. 2061-2070.

[22] F. ROTHE, *Global Solutions to Reaction-Diffusion Systems*, Springer-Verlag, Berlin, 1984.

[23] M. SERMANGE, *Study of a few equations based on the Hodgkin-Huxley model*, Math. Biosci., 36 (1977), pp. 45-60.

[24] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[25] L. TARTAR, *Quelques remarques sur les systèmes semi-linéaires*, in Proc. Coll. Iria-Novosibirsk, J. L. Lions and G. I. Marchouk, eds., Dunod, Paris, 1978.

[26] R. TEMAM, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, Berlin, New York, 1988.

[27] J. HALE, *Asymptotic behavior and dynamics in infinite dimensions*, in Nonlinear Differential Equations, J. Hale and P. Martinez-Amores, eds., Pitman, Boston, 1985.

[28] M. MARION, Attracteurs associés à des systèmes de réaction-diffusion partiellement dissipatifs, C.R. Acad. Sci. Paris Sér. I Math., 306 (1988), pp. 699-701.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 845-869, July 1989

005

LOCALIZED CLUSTER SOLUTIONS OF NONLINEAR DEGENERATE DIFFUSION EQUATIONS ARISING IN POPULATION DYNAMICS*

YUZO HOSONO† AND MASAYASU MIMURA‡

Abstract. The stationary problem of a nonlinear degenerate diffusion equation with nonlocal advection term including the aggregative mechanism is investigated. The spatially clustering phenomena of individuals in population dynamics is modeled. The existence of two kinds of stationary solutions with connected compact support—a large as well as a small cluster solution—is proved by the matched asymptotic expansion method and the geometric singular perturbation method, respectively.

Key words. nonlinear degenerate diffusion, nonlocal advection, aggregation, clustering, singular perturbation

AMS(MOS) subject classifications. 35K65, 34E15, 92A15

1. Introduction. In this paper, we will study stationary patterns of a one-dimensional phenomenological population model that describes the spatially clustering phenomena of individuals:

$$u_t = \varepsilon^2 (u^m)_{xx} - \varepsilon (K[u]u)_x + f(u), \qquad (x, t) \in R \times (0, \infty), \tag{1.1}$$

where $u(x, t)$ denotes the population density at position x and time t and the constants m and ε satisfy that $m > 1$ and $\varepsilon > 0$. Here, $K[u]$ is given by

$$K[u](x, t) = \lambda \left[\int_x^{x+r} u(y, t)\, dy - \int_{x-r}^{x} u(y, t)\, dy \right] \tag{1.2}$$

with parameters $\lambda > 0$ and r $(0 \leqq r \leqq \infty)$, and f is assumed to be a cubic-like function.

First, we briefly give the ecological interpretation on the model equation (1.1) with (1.2). The dispersal of individuals involves two processes. One is the density-dependent random movement that possesses "population pressure effect" (see Gurney and Nisbet [8], for instance). The other is the directed movement to cluster for oneself; that is, if the total population in the interval $(x-r, x)$ is less than that in $(x, x+r)$, that is, $\int_{x-r}^{x} u(y, t)\, dy < \int_x^{x+r} u(y, t)\, dy$, the individuals at position x move to the right and to the left if the inequality is reversed. Hence, (1.2) includes the aggregative mechanism of individuals through nonlocal interactions. $f(u)$ means the growth term of individuals. Let us show one specific form of $f(u)$ often used in mathematical ecology:

$$f_p(u) = \left(\alpha - \frac{u}{K} \right) u - \frac{ku}{1+u} P,$$

where α is the intrinsic growth rate of the individual. K is the carrying capacity, $ku/(1+u)$ is the predation rate of a predator with the maximum rate k, and P is the population density of the predator. If P remains constant and takes a suitable value, $f_p(u) = 0$ has three solutions, say O, A, and I as in Fig. 1. In this paper, we specify $f(u)$ as

$$f(u) = u(1-u)(u-a) \tag{1.3}$$

with a parameter a $(0 < a < 1)$, because it is a quite simple but suggestive nonlinearity.

* Received by the editors May 16, 1988; accepted for publication October 7, 1988.
† Institute of Computer Sciences, Kyoto Sangyo University, Kyoto, 603, Japan.
‡ Department of Mathematics, Hiroshima University, Hiroshima, 730, Japan.

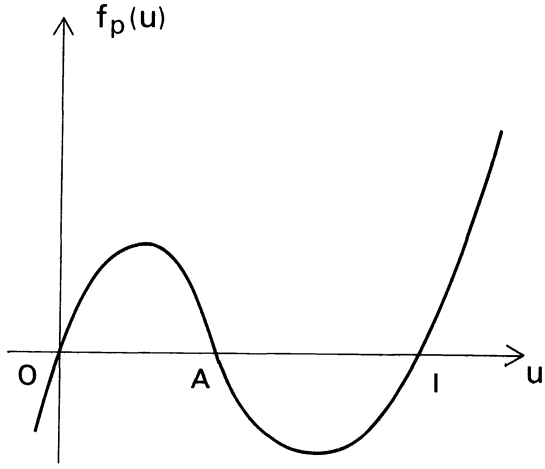


FIG. 1. *Functional form of* $f_p(u)$.

This kind of model has been discussed recently by several authors. For $f \equiv 0$ and $m > 1$, Nagai and Mimura [14]-[16] have fully discussed the case $r = \infty$ and have shown the existence of stationary solutions localized in a finite interval and their stability, and Ikeda [11] and Ikeda and Nagai [12] have investigated the case $r < +\infty$. See also Alt [1] for chemotactic models.

If $\lambda = 0$, that is, if there exists no aggregative effect, then (1.1) is reduced to the nonlinear degenerate diffusion equation with bistable kinetics (1.3):

$$u_t = \varepsilon^2 (u^m)_{xx} + f(u), \qquad (x, t) \in R \times (0, \infty). \tag{1.4}$$

Suppose that the initial data have compact support. Then, if $S(a, m) \equiv \int_0^1 f(u) u^{m-1}\, du < 0$, the solution $u(x, t)$ of (1.4) tends to zero uniformly in $x \in R$, while if $S(a, m) > 0$ and $u(x, 0) > a$ for all $x \in (-L, L)$ with sufficiently large positive L, $u(x, t)$ tends to 1 uniformly on any compact subset of R. Moreover, the support of solutions is compact for each $t > 0$ and monotone nondecreasing, that is, $\operatorname{supp}[u(\cdot, t_1)] \subset \operatorname{supp}[u(\cdot, t_2)]$ for any $0 \leqq t_1 < t_2$ (see Hosono [9]). This compactness property of support is one of the main features of nonlinear degenerate diffusion equations such as the porous media equation (see, for example, Aronson [2]).

The linear diffusion case (1.1) for $m = 1$ and $r = \infty$ has already been studied by Mimura, Terman, and Tsujikawa [13] for the case of bistable kinetics (1.3). The above results motivate us to consider (1.1) for $m > 1$ and $r < +\infty$, and show the existence of localized cluster solutions due to nonlocal advection (1.2).

In this paper, we assume that ε is sufficiently small, namely, the dispersal rate is much slower than the rate of dynamics, and consider the stationary problem of (1.1):

$$\begin{aligned} &\varepsilon^2 (u^m)_{xx} - \varepsilon (K[u]u)_x + f(u) = 0, \qquad x \in R, \\ &u(\pm\infty) = 0. \end{aligned} \tag{1.5}$$

We show that two types of solutions with compact connected support exist, depending on r and a: a large single cluster solution that has $O(1)$-length support and a small one that has $O(\varepsilon)$-length support (see Fig. 4). The numerical simulations suggest that the former is *stable* while the latter is *unstable*. Here, it should be noted that the degeneracy of diffusion at $u = 0$ requires the weak definition of solutions. A solution

of (1.5) is defined by a nonnegative function u on R satisfying the following: (i) u is bounded and continuous; (ii) u^m satisfies the integral identity

$$\int_{-\infty}^{+\infty} \{-[\varepsilon^2(u^m)_x - \varepsilon K[u]u]\varphi_x + f\varphi\}\, dx = 0$$

for all $\varphi \in C^1(R)$ such that $\varphi \geqq 0$ and vanishes for large values of $|x|$. To prove the existence of single cluster solutions, we employ two different approaches by exploiting the smallness of ε, that is, the singular perturbation method (the matched asymptotic expansion technique) for a large one and the shooting method used in [6] for a small one. In both cases, the degenerate character of the problem makes the proofs more complicated than for linear diffusion cases.

Due to the compact support of the stationary solutions, there also exist *multiple* cluster solutions. Figure 2a-f shows several cluster solutions; the number of clusters depends on the magnitude of the initial function in an appropriate sense. Such phenomena are drastically different from the linear version ($m = 1$).

2. Formulation of the problem and the main results. We consider a symmetric solution $u(x)$ at $x = 0$, with connected compact support, that is, it satisfies $u(x) = u(-x)$ for all $x \in R$ and $\operatorname{supp}[u] = [-\omega, \omega]$ with some $\omega > 0$. Then our problem is to find a solution $u(x; \varepsilon, \omega)$ of

$$\varepsilon^2(u^m)_{xx} - \varepsilon(K[u]u)_x + f(u) = 0, \qquad x \in I_\omega = (-\omega, \omega), \tag{2.1}$$

$$u = 0, \quad (u^m)_x = 0 \quad \text{at } x = I_{\pm\omega},$$

$$u(x) = u(-x) > 0 \quad \text{for } x \in I_\omega, \tag{2.2}$$

$$u(x) = 0 \quad \text{for } x \in R \setminus I_\omega,$$

where ω is a parameter to be determined. We note that solutions of (2.1), (2.2) become a solution of (1.5).

We will first construct a formal approximation to a large cluster solution of (2.1), (2.2). Set $\varepsilon = 0$ in (2.1). Then we have $f(u) = 0$, and define U_o by

$$U_o(x, \beta) = \begin{cases} 1, & 0 \leqq |x| < \beta, \\ 0, & |x| \geqq \beta, \end{cases} \tag{2.3}$$

for any fixed $\beta \in (0, \infty)$. We discuss the jump discontinuity of U_o only at $x = \beta$. Let us introduce the stretched variable $\xi = (x - \beta)/\varepsilon$. Then the equation in (2.1) is

$$(u^m)_{\xi\xi} - (K[u]u)_\xi + f(u) = 0,$$

where

$$K[u] = \lambda\left[\int_{\varepsilon\xi+\beta}^{\varepsilon\xi+\beta+r} u(y)\, dy - \int_{\varepsilon\xi+\beta-r}^{\varepsilon\xi+\beta} u(y)\, dy\right].$$

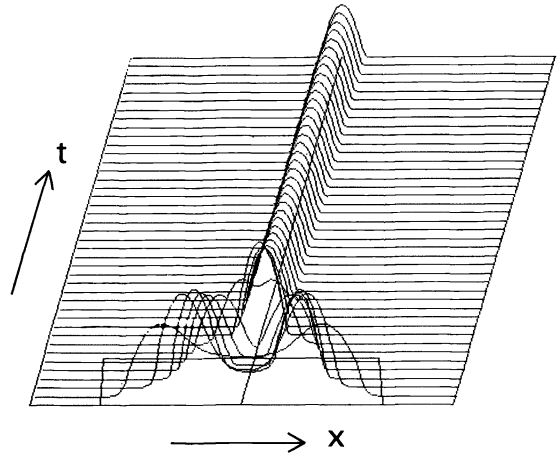
We may expect that the solution u is given by $U_o(x, \omega) + o(1)$ for $|x| \leqq \beta - \kappa(\varepsilon)$ and $|x| \geqq \beta + \kappa(\varepsilon)$ with some positive $\kappa(\varepsilon) = o(1)$, so that

$$K[u] = \lambda\left[\int_{\beta}^{\beta+r} U_o(y, \beta)\, dy - \int_{\beta-r}^{\beta} U_o(y, \beta)\, dy\right] + o(1) = -\lambda \min(r, 2\beta) + o(1).$$

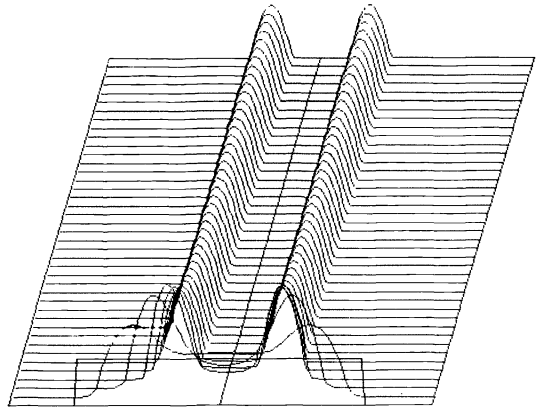
Setting $\varepsilon = 0$ and $k = \min(r, 2\beta)$, we have the inner problem:

$$(u^m)_{\xi\xi} + \lambda k u_\xi + f(u) = 0, \qquad \xi \in R, \tag{2.4}$$

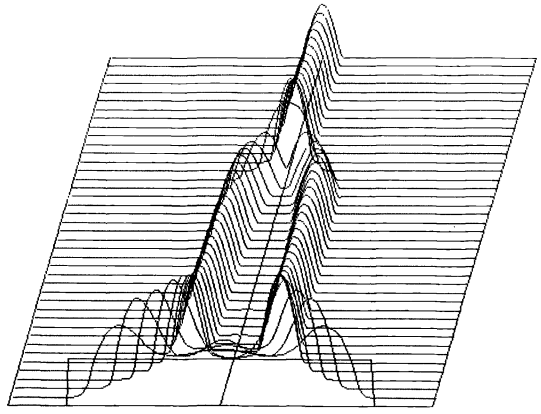
$$u(-\infty) = 1, \quad u(\beta_1) = (u^m)_x(\beta_1) = 0, \quad u(0) = \alpha \in [0, 1),$$



(a) $L = 10$.

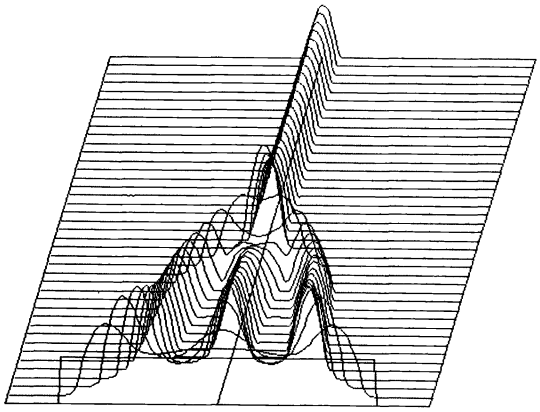


(b) $L = 11$.

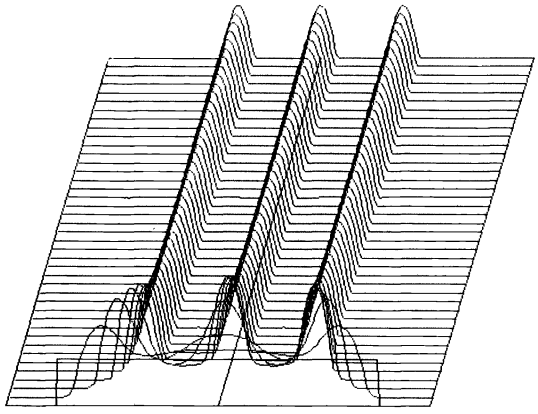


(c) $L = 13$.

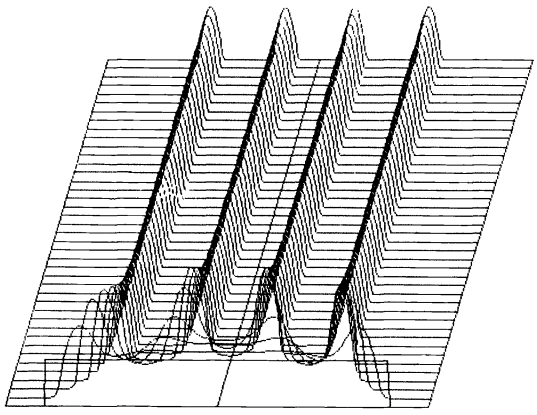
FIG. 2. *Time development of initial functions into multiple cluster solutions for the initial data* $u_0(x) = 1$ *for* $|x| \leqq L$ *and* $u_0(x) = 0$ *for* $|x| > L$, *where* $\varepsilon = \lambda = 1.0$, $a = 0.2$, $m = 2$, *and* $r = 5.0$.



(d) $L = 15$.



(e) $L = 16$.



(f) $L = 2$.

FIG. 2—continued

where the last condition is imposed to fix the translation invariance and β_1 is an unknown to be determined. For this problem, it is shown in [7] that there exist a unique $c^*(a)$ and a unique positive ω_1 such that (2.4) has a unique monotone decreasing solution $U_i(\xi, \beta_1)$ in the weak sense if and only if $k = c^*(a)/\lambda$ and $\beta_1 = \omega_1$. Moreover, $c^*(a)$ is a strictly monotone decreasing function of $a \in (0, 1)$ satisfying $c^*(a) = 0$ for $a = a^* = (m+1)/(m+3)$ and $c^*(a) \gtreqless 0$ according to $S(a, m) \gtreqless 0$ (see [7]). Therefore,

if $a \in (0, a^*)$ and r satisfies $r \geqq 2\omega_0$ with $\omega_0 = c^*(a)/2\lambda > 0$, we can choose β as $\beta = \omega_0$ so that $k = \min(r, 2\beta)$ attains the value $c^*(a)/\lambda$. Thus, we have obtained the lowest-order outer and inner solutions as $U_o(x, \omega_0)$ and $U_i(\xi, \omega_1) = U_i((x-\omega_0)/\varepsilon, \omega_1)$, respectively.

Next, we consider a small cluster solution. Taking its outer solution as identically zero, we may construct only the inner solution. Put $\omega = \varepsilon\sigma$ and $\eta = x/\varepsilon$, and rewrite (2.1) as follows:

$$(u^m)_{\eta\eta} - \varepsilon(\tilde{K}[u]u)_\eta + f(u) = 0, \qquad \eta \in J = (-\sigma, \sigma), \tag{2.5}$$

$$u = (u^m)_\eta = 0 \quad \text{at } \eta = \pm\sigma,$$

$$u(\eta) = u(-\eta) > 0 \quad \text{for } \eta \in J, \tag{2.6}$$

$$u(\eta) \equiv 0 \quad \text{for } \eta \in R \backslash J,$$

where

$$\tilde{K}[u] = \lambda\left[\int_\eta^{\eta+r/\varepsilon} u(\eta')\,d\eta' - \int_{\eta-r/\varepsilon}^{\eta} u(\eta')\,d\eta'\right].$$

Setting $\varepsilon = 0$ in (2.5), we have

$$(u^m)_{\eta\eta} + f(u) = 0, \qquad \eta \in J, \tag{2.7}$$

which can be solved under the conditions (2.6) (see Lemma 4.1 and its remark). We see that, for each $a \in (0, a^*)$, there is a unique $\sigma_0 > 0$ such that a solution $u_i(\eta, \sigma_0)$ of (2.7) uniquely exists for $\sigma = \sigma_0$ and is monotone decreasing for $\eta \geqq 0$. Moreover, σ_0 is a strictly monotone increasing function of a satisfying $\lim_{a\to a^*} \sigma_0 = +\infty$ and $\lim_{a\to 0} \sigma_0 = 0$. This is the lowest-order approximation to a small solution. Now we state our theorem.

THEOREM 2.1 (Existence of single cluster solutions). *Assume that $0 < a < a^*$ and let δ be any small positive constant. Then, we have the following*:

(A) *For each fixed $r \geqq 2\omega_0 + \delta$, there exists an $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$, (2.1), (2.2) has a large single (symmetric) cluster solution $U_l(x, \varepsilon)$ satisfying that for any small $\kappa > 0$,*

$$\lim_{\varepsilon\to 0} U_l(x, \varepsilon) = 1 \quad \textit{uniformly in } x \in (-\omega_0 + \kappa, \omega_0 - \kappa) \tag{2.8}$$

and

$$\operatorname{supp}[U_l] = [-\omega_0 - \tilde{\omega}_1(\varepsilon), \omega_0 + \tilde{\omega}_1(\varepsilon)], \tag{2.9}$$

where $\tilde{\omega}_1(\varepsilon)$ is a positive function of ε satisfying $\lim_{\varepsilon\to 0} \tilde{\omega}_1(\varepsilon) = 0$.

(B) *For each fixed $r \geqq \delta > 0$, there exists an $\varepsilon_1 > 0$ such that for any $\varepsilon \in (0, \varepsilon_1)$, (2.1), (2.2) has a small single (symmetric) cluster solution $U_s(x, \varepsilon)$ satisfying*

$$\lim_{\varepsilon\to 0} U_s(\varepsilon\eta, \varepsilon) = u_i(\eta; \sigma_0) \quad \textit{uniformly in } \eta = \frac{x}{\varepsilon} \in R \tag{2.10}$$

and

$$\operatorname{supp}[U_s] = [-\varepsilon\sigma_\varepsilon, \varepsilon\sigma_\varepsilon], \tag{2.11}$$

where σ_ε is some positive function of ε satisfying $\lim_{\varepsilon\to 0} \sigma_\varepsilon = \sigma_0$.

Let u_1 and u_2 be any two solutions of (1.5) satisfying that the distance between $\operatorname{supp}[u_1]$ and $\operatorname{supp}[u_2]$ is not less than r; that is, $\operatorname{dist}(\operatorname{supp}[u_1], \operatorname{supp}[u_2]) \geqq r$. Then, it is easily seen that $u_1 + u_2$ becomes a solution of (1.5), so that our theorem directly gives the following corollary.

COROLLARY (Existence of multiple cluster solutions). *Let* $\{x_i\}_{i\in\Lambda}$ (Λ: *an index set*) *be a finite or countable sequence of real numbers and* Λ_1, Λ_2 *be index sets satisfying* $\Lambda=\Lambda_1\cup\Lambda_2$, $\Lambda_1\cap\Lambda_2=\phi$, *and denote* $U_i=U_l(x-x_i,\varepsilon)$ *for* $i\in\Lambda_1$ *and* $u_j=U_s(x-x_j,\varepsilon)$ *for* $j\in\Lambda_2$. *If* $\operatorname{dist}(\operatorname{supp}[u],\operatorname{supp}[\tilde{u}])\geqq r$ *for any* $u,\hat{u}\in\{U_i\}_{i\in\Lambda_1}\cup\{u_j\}_{j\in\Lambda_2}$ *such that* $u\neq\tilde{u}$, *then* (1.5) *has also a solution* $u(x,\varepsilon)$ *represented by*

$$u(x,\varepsilon)=\sum_{i\in\Lambda_1}U_i+\sum_{j\in\Lambda_2}u_j \tag{2.12}$$

(*see Fig.* 2).

Remark 1. If $a\geqq a^*$, then $c^*(a)\leqq 0$, so that there exists no solution for (2.4) since $k=\min(r,2\beta)>0$. Also (2.7) and (2.6) have no solution for $a\geqq a^*$. By this result and the complementary numerical simulations, we may conjecture that (2.1) and (2.2) have no spatially heterogeneous steady state solution when $a\geqq a^*$.

Remark 2. Let r be fixed in $(0,2\omega_0^*)$ with $\omega_0^*\equiv\lim_{a\to 0}c^*(a)/2\lambda$ and define a_r^* by the relation $r=c^*(a_r^*)/\lambda$. Then $k\leqq r<2\omega_0^*$ for all $\beta>0$, which implies that there exists no solution of (2.4) for $a\in[0,a_r^*]$. Hence, for small $\varepsilon>0$, we can expect that there is some $a_r(\varepsilon)$ satisfying $\lim_{\varepsilon\to 0}a_r(\varepsilon)=a_r^*$ such that for $0\leqq a\leqq a_r(\varepsilon)$ there exists no large cluster solution and any solution expands into R, as in Fig. 3, or decays to zero when initial data have compact support.

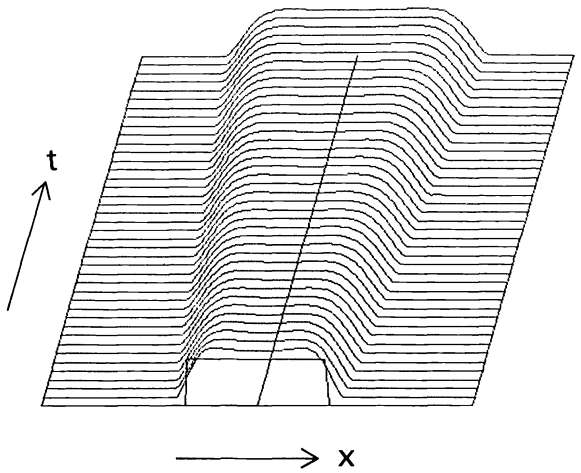


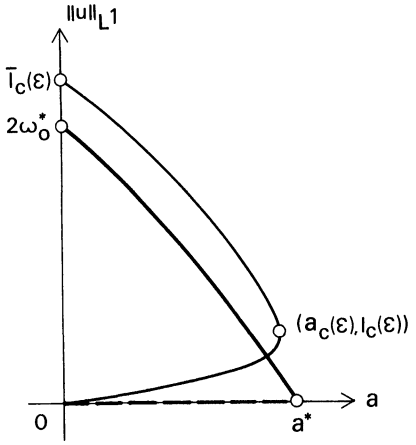
FIG. 3. *An expanding wave solution for* $r=1.5<2\omega_0$, $\varepsilon=\lambda=1.0$, $a=0.1$, *and* $m=2$.

These results suggest the global branch of single cluster solutions in $(a,\|u\|_{L^1})$-space as $\varepsilon\to 0$, where $\|u\|_{L^1}$ denotes the L^1-norm of a solution u. For each fixed $r\geqq 0$, we consider the branch $\|u\|_{L^1}$ as a function of a. For $r\geqq 2\omega_0^*$, the solution branch is as in Fig. 4(a) where there are large (bold line) as well as small (bold broken line) singular solutions (as $\varepsilon\to 0$ for $0<a<a^*$), and for $0<r<2\omega_0^*$, the branch is as in Fig. 4(b), where there is a large singular solution only for $a_r^*<a<a^*$. For $r=0$, there is only a small singular solution branch for $0<a<a^*$. From the result of the singular case, we may conjecture the following global structure of single cluster solutions for $\varepsilon>0$. We first note that there are two numbers $r_1(\varepsilon)$ and $r_2(\varepsilon)$ satisfying $\lim_{\varepsilon\to 0}r_1(\varepsilon)=0$ and $\lim_{\varepsilon\to 0}r_2(\varepsilon)=2\omega_0^*$, such that for each fixed $r\geqq r_2(\varepsilon)$, the (unstable) small solution branch starting from the origin goes right up to a limit point, say $(a_c(\varepsilon),l_c(\varepsilon))$ and through this point, it goes back to the left with the recovery of stability. Then it arrives at a point on $a=0$, say $(0,\bar{l}_c(\varepsilon))$. Here it should be noted that $a_c(\varepsilon)\to a^*$, $l_c(\varepsilon)\to 0$, and $\bar{l}_c(\varepsilon)\to 2\omega_0^*$ as $\varepsilon\to 0$. For a fixed $r\in(r_1(\varepsilon),r_2(\varepsilon))$, there

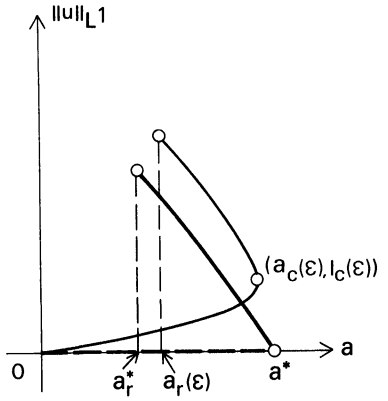
is a large solution only for $a_r(\varepsilon) < a < a_c(\varepsilon)$, and for $r \in (0, r_1^*(\varepsilon))$ there exists no large cluster solution. The cases when $\varepsilon > 0$ are depicted in Fig. 4a-c by solid lines.

We devote the remaining part of this paper to the proof of our theorem.

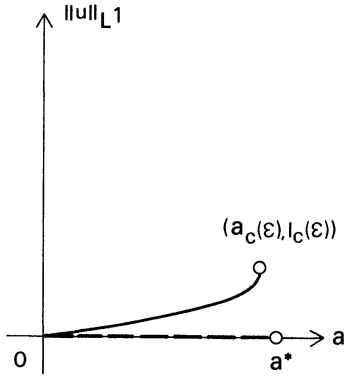
3. Existence of a large single cluster solution. Set $v(x) = \int_0^x u(y)\,dy$ and assume $r \geqq 2\omega$. This implies that $v(x+r) = -v(x-r) = \int_0^\omega u(y)\,dy$ for $|x| \leqq \omega$ and $u(x) \equiv 0$ for $|x| > \omega$, so that we see $K[u]u = -\lambda 2vu$. Here, we introduce the new variable $q = u^{m-1}$.



(a) $r \geqq 2\omega_0$.



(b) $\delta < r < 2\omega_0$.



(c) $r = O(\varepsilon)$ *and* $r \geqq \varepsilon\sigma_0$.

FIG. 4. *The global solution branch in* $(a, \|u\|_{L^1})$*-space for* $\varepsilon = 0$ *and* $\varepsilon > 0$.

Then our problem is to find a solution (q, v) and ω of

$$N(q, v) \equiv \begin{pmatrix} \varepsilon^2 q q_{xx} + \varepsilon^2 \mu q_x^2 + 2\varepsilon \dfrac{\lambda}{m} v q_x + F_\varepsilon(q) \\ v_x - q^\mu \end{pmatrix} = 0 \quad \text{for } x \in I_\omega = (0, \omega) \tag{3.1}$$

and

$$\begin{aligned} &q_x(0) = q(\omega) = (q^{\mu m})_x(\omega) = 0, \\ &q(x) > 0 \quad \text{for } x \in I_\omega, \qquad q(x) = 0 \quad \text{for } \bar{R}_\omega = [\omega, +\infty), \\ &v(0) = 0, \end{aligned} \tag{3.2}$$

where $\mu = 1/(m-1)$ and $F_\varepsilon(q) = (1/m\mu)[q^{1-\mu} f(q^\mu) + 2\varepsilon\lambda q^{1+\mu}]$. To solve this problem, we employ the singular perturbation method with the patching argument (see, for example, [7]). Hence, we split it into the following two problems. Let α be a fixed constant satisfying $a^{m-1} < \alpha < 1$, and let β and γ be parameters in some neighborhood $J_\delta = (\omega_0 - \delta, \omega_0 + \delta)$ of ω_0 with a small $\delta > 0$. The first problem is to find a solution $(\tilde{q}, \tilde{v})$ of

$$\begin{aligned} &N(q, v) = 0, \qquad x \in I_\beta = (0, \beta), \\ &q_x(0) = 0, \quad q(\beta) = \alpha, \quad v(0) = 0. \end{aligned} \tag{3.3}$$

The second is to find a solution $(\hat{q}, \hat{v})$ and ω of

$$\begin{aligned} &N(q, v) = 0, \qquad x \in R_\beta = (\beta, +\infty), \\ &q(\beta) = \alpha, \quad q(\omega) = (q^{m\mu})_x(\omega) = 0, \quad q(x) > 0 \quad \text{for } x \in [\beta, \omega), \\ &v(\beta) = \gamma, \end{aligned} \tag{3.4}$$

which has the degenerate character passed from (3.1), (3.2). We first discuss (3.4) and later (3.3).

Throughout this paper, we will use the following Banach spaces: For $I = (0, 1)$ and $J = R_+ = (0, +\infty)$:

$$C_0^1(I) = \left\{ u \middle| u \in C^1(I), \sup_{x \in I} \left(|u(x)| + \left| \frac{du}{dx}(x) \right| \right) < +\infty, u(0) = 0 \right\},$$

$$C_{\varepsilon,0}^2(I) = \left\{ u \middle| u \in C^2(I), \sup_{x \in I} \sum_{i=0}^{2} \left| \left(\varepsilon \frac{d}{dx} \right)^i u(x) \right| < +\infty, u_x(0) = u(1) = 0 \right\},$$

$$X_\rho^p(J) = \left\{ u \middle| u \in C^p(J), \sup_{x \in J} e^{\rho x} \sum_{i=0}^{p} \left| \left(\frac{d}{dx} \right)^i u(x) \right| < +\infty \right\},$$

$$X_{\rho,0}^p(J) = \{ u | u \in X_\rho^p(J), u(0) = 0 \},$$

$$X_{,\rho}^1(J) = \{ u | u \in X_{0,0}^0(J), u_x \in X_\rho^0(J) \},$$

$$X^0(J) = X_0^0(J).$$

We often write $C_0^1(I)$ as simply C_0^1, and do other spaces in a similar way. C_j $(j = 0, 1, 2, \cdots)$ always denote positive constants independent of ε.

3.1. Construction of solutions of the degenerate problem. Set $\xi = (x - \beta)/\varepsilon$ and $p = q_\xi$, and introduce the new independent variable ζ by

$$\frac{d\xi}{d\zeta} = q, \qquad \xi(0) = 0. \tag{3.5}$$

Then, (3.4) is written as

$$(3.6)_\varepsilon \qquad \begin{aligned} q_\zeta &= pq, \\ p_\zeta &= -\mu p^2 - \frac{2\lambda}{m} vp - F_\varepsilon(q), \qquad \zeta \in R_+ = (0, +\infty), \\ v_\zeta &= \varepsilon q^{1+\mu}, \end{aligned}$$

and

$$(3.7) \qquad \begin{aligned} &q(0) = \alpha, \quad q(+\infty) = 0, \quad p(\zeta) < 0 \quad \text{for } \zeta \in R_+, \\ &v(0) = \gamma. \end{aligned}$$

For $\varepsilon = 0$, assuming that δ is sufficiently small, we already have the result that for $\gamma \in J_\delta$, there exists a unique solution $y_0(\zeta; \gamma) = (q_0(\zeta: \gamma), p_0(\zeta; \gamma), v_0(\zeta; \gamma))$ of $(3.6)_\varepsilon$, (3.7) which satisfies that $q_0, p_0 + 2\lambda\gamma/m\mu \in X^1_{-\rho_-}(R_+)$ with $\rho_- = -2\lambda\gamma/m\mu$, $v_0(\zeta; \gamma) \equiv \gamma$, and $p_0(0, \gamma)$ is monotone decreasing with respect to γ (see [10, Lemma 5.6]). For $(3.6)_\varepsilon$, (3.7) with a small $\varepsilon > 0$, we will look for a solution $\hat{y}$ of $(3.6)_\varepsilon$, (3.7) in the form $\hat{y} = y_0 + y$, where $y = {}^t(q_1, p_1, v_1)$. Let $y_1 = {}^t(q_1, p_1)$. Then, $(3.6)_\varepsilon$ is written as

$$(3.8) \qquad G(y, \varepsilon) = \begin{pmatrix} Q_1 y_1 - g_1(y, \varepsilon) \\ \dfrac{d}{d\zeta} v_1 - g_2(q_1, \varepsilon) \end{pmatrix} = 0,$$

where

$$Q_1 = \frac{d}{d\zeta} - B(\zeta), \qquad B(\zeta) = \begin{pmatrix} p_0 & q_0 \\ -F_0'(q_0) & -2(\mu p_0 + \lambda\gamma/m) \end{pmatrix}$$

with $F_0(q) = q^{1-\mu} f(q^\mu)/m\mu$,

$$g_1(y, \varepsilon) = \begin{pmatrix} p_1 q_1 \\ -[F_0(q_0 + q_1) - F_0(q_0) - F_0'(q_0) q_1] - \mu p_1^2 - \dfrac{2\lambda}{m}(p_0 + p_1) r - \dfrac{2\lambda\varepsilon}{m\mu}(q_0 + q_1)^{1+\mu} \end{pmatrix}$$

and $g_2(q_1, \varepsilon) = \varepsilon(q_0 + q_1)^{1+\mu}$. The boundary condition is

$$(3.9) \qquad q_1(0) = q_1(+\infty) = 0, \qquad v_1(0) = 0.$$

From (3.8), (3.9), for a small $\varepsilon_0 > 0$ and some $\rho \in (0, \rho_\delta]$ with $\rho_\delta = -\rho_- - \delta$ $(\delta > 0)$, we define a nonlinear mapping

$$G(y, \varepsilon): Z^1_{\rho,0} \times [0, \varepsilon_0) \equiv X^1_{\rho,0} \times X^1_0 \times X^1_{,\rho} \times [0, \varepsilon_0) \to Z^0_\rho \equiv X^0_\rho \times X^0 \times X^0.$$

Then our problem is to find a solution $y(\varepsilon) \in Z^1_{\rho,0}$ of $G(y, \varepsilon) = 0$. Let us further rewrite this as follows. Integrating the second equation in (3.8), we see that if $q_1 \in X^0_{\rho,0}$, then

$$(3.10) \qquad v_1(\zeta) = \varepsilon \int_0^\zeta [q_0(\zeta') + q_1(\zeta')]^{1+\mu} \, d\zeta' = \varepsilon R_1(q_1) \in X^1_{\rho,0}$$

for any $\rho \in (0, \rho_\delta]$. Therefore, (3.8), (3.9) is reduced to

$$(3.11) \qquad Q_1 y_1 = g_1(y_1, \varepsilon R(q_1), \varepsilon) = h_1(y_1) + \varepsilon h_2(y_1),$$

where

$$h_1(y_1) = \begin{pmatrix} h_{11}(y_1) \\ h_{12}(y_1) \end{pmatrix} = \begin{pmatrix} p_1 q_1 \\ -[F_0(q_0 + q_1) - F_0(q_0) - F_0'(q_0) q_1] - \mu p_1^2 \end{pmatrix}$$

and

$$h_2(y_1)=\begin{pmatrix}0\\ h_{22}(y_1)\end{pmatrix}$$

$$=\begin{pmatrix}0\\ -\dfrac{2\lambda}{m}(p_0+p_1)R(q_1)-\dfrac{2\lambda}{m\mu}(q_0+q_1)^{1+\mu}\end{pmatrix}.$$

For the linear mapping Q_1, we have Lemma 3.1.

LEMMA 3.1. *Let δ be sufficiently small. Then for each $\rho\in(0,\rho_\delta]$, the map Q_1: $Y_\rho^1\equiv X_{\rho,0}^1\times X_0^1\to Y_\rho\equiv X_\rho^0\times X^0$ has a bounded inverse Q_1^{-1}.*

Proof. See the Appendix for the proof.

Applying Q_1^{-1} to (3.11), we have the equation

$$y_1=Q^{-1}g_1(y_1,2R(q_1),\varepsilon)\equiv\mathcal{F}(y_1,\varepsilon). \tag{3.12}$$

The nonlinear operator $\mathcal{F}$: $Y_\rho\times[0,\varepsilon_0)\to Y_\rho^1\subset Y_\rho$ satisfies the following lemma.

LEMMA 3.2. *Let ρ be fixed in $(0,\rho_\delta]$ and $\gamma\in J_\delta$. Then there exist a small closed ball $B_\kappa(0)=\{y_1\,|\,\|y_1\|_{Y_\rho}\leqq\kappa\}$ and a small constant $\varepsilon_0>0$ such that $\mathcal{F}$ is a uniform contraction on $B_\kappa(0)$ and continuous from $B_\kappa(0)\times[0,\varepsilon_0)$ into $B_\kappa(0)$.*

Proof. See the Appendix for the proof.

Thus, we can apply the uniform contraction principle [4, Thm. 2.2] to (3.12) and obtain a solution $y_1(\varepsilon)=(q_1(\zeta;\varepsilon,\gamma),\ p_1(\zeta;\varepsilon,\gamma))\in Y_\rho$ satisfying $y_1(0)=0$. Moreover, (3.10), (3.12) shows that $y(\zeta;\varepsilon,\gamma)=(q_1(\zeta;\varepsilon,\gamma),\ p_1(\zeta;\varepsilon,\gamma),\ v_1(\zeta;\varepsilon,\gamma))\in Z_\rho^1$ with $v_1(\zeta;\varepsilon,\gamma)=\varepsilon R_1(q_1(\zeta;\varepsilon,\gamma))$ and it satisfies $\lim_{\varepsilon\to0}\|y(\zeta;\varepsilon,\gamma)\|_{Z_\rho^1}=0$. Hence we have a solution $y_0(\zeta;\gamma)+y(\zeta;\varepsilon,\gamma)$ of $(3.6)_\varepsilon$, (3.7). Note that the above argument is valid uniformly in $\gamma\in J_\delta$ for a small $\delta>0$, so that y is uniformly continuous in ε and γ.

Next, we examine relation (3.5). Let $\phi_0(\zeta)=\int_0^\zeta q_0(\zeta';\gamma)\,d\zeta'$ and $\phi_\varepsilon(\zeta)=\int_0^\zeta[q_0(\zeta';\gamma)+q_1(\zeta';\varepsilon,\gamma)]\,d\zeta'$. Then

$$\sup_{\zeta\in R_+}|\phi_\varepsilon(\zeta)-\phi_0(\zeta)|\leqq\sup_{\zeta\in R_+}\int_0^\zeta|q_1(\zeta';\varepsilon,\gamma)|\,d\zeta'\leqq\|q_1\|_{X_\rho^0}\int_0^\infty e^{-\rho\zeta'}\,d\zeta'\to0$$

as $\varepsilon\to0$ uniformly in $\gamma\in I_\delta$. Set $\lim_{\zeta\to\infty}\phi_\varepsilon(\zeta)=\omega_\varepsilon(\gamma)<+\infty$ for $\varepsilon\in[0,\varepsilon_0)$. Since $d\phi_\varepsilon/d\zeta>0$ for $\zeta\in[0,\infty)$, ϕ_ε is a diffeomorphism from $[0,\infty)$ onto $[0,\omega_0)$, so that we have an inverse function ϕ_ε^{-1} of ϕ_ε. Using this, we define a pair of functions $(\hat q,\hat v)$ by

$$\hat q(\xi;\varepsilon,\gamma)=\begin{cases}\hat q_0(\xi;\gamma)+q_1(\phi_\varepsilon^{-1}(\xi);\varepsilon,\gamma), & 0<\xi<\omega_\varepsilon,\\ 0, & \omega_\varepsilon\leqq\xi<+\infty,\end{cases}$$

$$\hat v(\xi;\varepsilon,\gamma)=\gamma+\varepsilon\int_0^\xi[\hat q(\xi';\varepsilon,\gamma)]^\mu\,d\xi'$$

with $\xi=(x-\beta)/\varepsilon$, where $\hat q_0(\xi;\gamma)=q_0(\phi_\varepsilon^{-1}(\xi);\gamma)$. This is a solution of (3.4). Thus we have Theorem 3.1.

THEOREM 3.1. *Let $\delta>0$ be sufficiently small. Then, for $\gamma\in I_\delta$ there exists a small $\varepsilon_0>0$ such that for $\varepsilon\in(0,\varepsilon_0)$, (3.4) has a solution $(\hat q(\xi;\varepsilon,\gamma),\ \hat v(\xi;\varepsilon,\gamma))$ with $\xi=(x-\beta)/\varepsilon$ satisfying $\hat q>0$ for $0\leqq\xi<\omega_\varepsilon(\gamma)<+\infty$, $\hat q=0$ for $\xi\geqq\omega_\varepsilon(\gamma)$, and $\lim_{\varepsilon\to0}(d/d\xi)\hat q(0;\varepsilon,\gamma)=(d/d\xi)\hat q_0(0;\gamma)$ uniformly in γ. Moreover, $(\hat q,\hat v)$ is uniformly continuous in ε and γ relative to the norm of $C^2(I_{\kappa'})\times C^2(I_{\kappa'})$ with $I_{\kappa'}=[0,\omega_0-\kappa']$ for any $\kappa'>0$ and $(d/d\xi)\hat q_0(0;\gamma)$ is strictly monotone decreasing with respect to γ.*

Remark 3.1. $\hat{q}_0(\xi;\gamma)$ is a unique monotone solution of

$$
\begin{aligned}
&qq_{\xi\xi}+\mu q_\xi^2+2\frac{\lambda\gamma}{m}q_\xi+F_0(q)=0, \qquad \xi\in(0,\omega_0),\\
&q(0)=\alpha, \qquad q(\omega_0)=0.
\end{aligned}
\tag{3.13}
$$

3.2. Construction of solutions of the nondegenerate problem. To make the parameter dependency of solutions clear, we normalize the interval by $z=x/\beta$ and set $\sigma=\varepsilon/\beta$. Then, (3.3) is

$$
\begin{aligned}
&\tilde{N}(q,v)\equiv\begin{pmatrix}\sigma^2qq_{zz}+\sigma^2\mu q_z^2+2\sigma\dfrac{\lambda}{m}vq_z+F_\sigma(q)\\ v_z-\beta q^\mu\end{pmatrix}=0,\\
&q_z(0)=0, \quad q(1)=\alpha, \quad v(0)=0,
\end{aligned}
\qquad z\in I=(0,1),
\tag{3.14}
$$

where $F_\sigma(q)=[q^{1-\mu}f(q^\mu)+2\lambda\sigma\beta q^{1+\mu}]/m\mu$. For sufficiently small $\sigma>0$, the standard singular perturbation method gives the lowest-order approximation as

$$
\tilde{q}_0(z;\sigma,\beta)=1+\theta(z-1)[q_0(\eta;\beta)-1],
$$

$$
\tilde{v}_0(z;\sigma,\beta)=\beta\int_0^z\tilde{q}_0(z';\sigma,\beta)^\mu\,dz',
$$

where $\eta=(z-1)/\sigma$ and $q_0(\eta,\beta)$ is a unique monotone decreasing solution of

$$
\begin{aligned}
&qq_{\eta\eta}+\mu q_\eta^2+2\frac{\lambda\beta}{m}q_\eta+F_0(q)=0, \qquad \eta\in R_-=(-\infty,0),\\
&q(-\infty)=1, \qquad q(0)=\alpha\in(0,1),
\end{aligned}
\tag{3.15}
$$

and $\theta(z)$ is a C^∞ cutoff function such that $0\leqq\theta(z)\leqq1$, $\theta(z)=1$ for $|z|\leqq\frac{1}{4}$ and $\theta(z)=0$ for $|z|\geqq\frac{1}{2}$. Note here that $|q_0(\eta;\beta)-1|$ decays exponentially with exponent $[-\lambda\beta+((\lambda\beta)^2-F_0'(1))^{1/2}]/m$ and $(d/d\eta)q_0(0;\beta)$ is strictly monotone increasing with respect to β. Using this approximation, we seek a solution in the form $(q,v)=(\tilde{q}_0,\tilde{v}_0)+(q_1,v_1)$.

For small positive σ_0 and δ, let us define a nonlinear mapping

$$
S(t,\sigma,\beta)=\tilde{N}(\tilde{q}_0+q_1,\tilde{v}_0+v_1)\colon Y_\sigma^2\times(0,\sigma_0)\times J_\delta\to Y^0,
$$

where $t=(q_1,v_1)$, $Y_\sigma^2=C_{\sigma,0}^2\times C_0^1$, and $Y^0=C^0\times C^0$. The Fréchet derivative of S,

$$
S_t(t,\sigma,\beta)=\begin{pmatrix}S_{11}&S_{12}\\S_{21}&S_{22}\end{pmatrix}
$$

is given by

$$
S_{11}(t,\sigma,\beta)=\sigma^2q\left(\frac{d}{dz}\right)^2+2\sigma\left(\mu\sigma q_z+\frac{\lambda}{m}v\right)\frac{d}{dz}+[\sigma^2q_{zz}+F'(q)],
$$

$$
S_{12}(t,\sigma,\beta)=2\sigma\frac{\lambda}{m}q_z, \quad S_{21}(t,\sigma,\beta)=-\mu\beta q^{\mu-1}, \quad S_{22}=\frac{d}{dz}.
$$

Let $B_\kappa(0)=\{t\,|\,\|t\|_{Y_\rho^2}\leqq\kappa\}$. It follows from this expression that

$$
\|S_t(t_1,\alpha,\beta)-S_t(t_2,\alpha,\beta)\|_{Y_\sigma^2\to Y^0}\leqq K_1\|t_1-t_2\|_{Y_\sigma^2}
$$

for $t_1,t_2\in B_\kappa(0)$, $\sigma\in(0,\sigma_0)$, and $\beta\in J_\delta$, where K_1 is some constant. We show the uniform invertibility of S_t.

LEMMA 3.3. *There exist small positive σ_0 and δ such that for any $\sigma \in (0, \sigma_0)$, $\alpha \in (1-\delta, 1)$, and $\beta \in J_\delta$, $S_t(0; \sigma, \beta)$ has an inverse S_t^{-1} satisfying $\|S_t^{-1}\|_{Y^0 \to Y^2_\sigma} \leqq K_s$, where K_s is a positive constant independent of σ and β.*

Proof. We first consider

$$S_{11}(0, \sigma, \beta) = \sigma^2 a_0 \left(\frac{d}{dz}\right)^2 + \sigma a_1 \frac{d}{dz} + a_2 : C^2_{\sigma,0} \to C^0,$$

where $a_0 = \tilde{q}_0$, $a_1 = 2(\mu \tilde{q}_{0z} + (\lambda/m)\tilde{v}_0)$, and $a_2 = \sigma^2 \tilde{q}_{0zz} + F'_\sigma(\tilde{q}_0)$. Since $\tilde{q}_0$ satisfies $\sup_{z \in I} |(\sigma(d/dz))^i [\tilde{q}_0 - 1]| \leqq C_i(1-\alpha)$ $(i = 0, 1, 2)$ with $C_0 = 1$, we easily find that $a_0 \geqq \alpha > 0$ and $a_2 \leqq C_2(1-\alpha) + F'_0(1) + F'_0(\tilde{q}_0) - F'_0(1) + 2\sigma\beta(\lambda/m)\mu \tilde{q}_0^{\mu-1} < 0$ for $\alpha \in (1-\delta, 1)$, $\beta \in J_\delta$, $\sigma \in (0, \sigma_0)$ if we choose δ and σ_0 sufficiently small. Therefore, the maximum principle for usual two-point boundary value problems shows that S_{11} has a uniformly bounded inverse $S_{11}^{-1}: C^0 \to C^2_{\sigma,0}$ satisfying $\|S_{11}^{-1}\| \leqq K_1$, where K_1 denotes a positive constant independent of σ, α, and β. It is obvious that $S_{22}(0, \sigma, \beta): C^1_0 \to C^0$ has an inverse S_{22}^{-1} satisfying $\|S_{22}^{-1}\|_{C^0 \to C^1_0} = 2$. S_{12} and S_{21} are multiplication operators satisfying

$$|S_{12}(0, \sigma, \beta)| \leqq \sup_{z \in I} \left| \frac{2\lambda}{m} \sigma \frac{d}{dz} \tilde{q}_0 \right| < \frac{2\lambda}{m} C_1(1-\alpha)$$

and

$$|S_{21}(0, \sigma, \beta)| \leqq \sup_{z \in I} |\mu\beta \tilde{q}_0^{\mu-1}| \leqq \mu\beta \max(1, (1-\alpha)^{\mu-1}) < C_3.$$

These estimates assure the uniform invertibility of S_t. Actually, consider $S_t(0, \sigma, \beta)t = k$ for $k = (k_1, k_2) \in Y^0$. Then we can easily rewrite it as follows:

$$q_1 = S_{11}^{-1}(k_1 - S_{12}S_{22}^{-1}k_2) - S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}q_1,$$
$$v_1 = S_{22}^{-1}k_2 - S_{22}^{-1}S_{21}q_1.$$

Since $\|S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}\|_{C^2_{\sigma,0} \to C^2_{\sigma,0}} \leqq (4\lambda/m)C_1C_3K_1(1-\alpha) \leqq \frac{1}{2}$ if we choose α as $\alpha \geqq 1 - m/8\lambda C_1C_3K_1$, it follows from the above equations that $\|q_1\|_{C^2_{\sigma,0}} \leqq C_4\|k\|_{Y^0}$ and $\|v_1\|_{C^1_0} \leqq C_5\|k\|_{Y^0}$. Here all constants C_i are independent of σ, α, and β. This completes the proof.

LEMMA 3.4. $\lim_{\sigma \to 0} \|S(0, \sigma, \beta)\|_{Y^0} = 0$ *uniformly in $\beta \in J_\delta$ for small $\delta > 0$.*

Proof. This lemma can be proved by the standard matched asymptotic technique (see, for example, [10]), so we omit the proof.

Now Lemmas 3.2 and 3.3 enable us to apply the implicit function theorem to $S(t, \sigma, \beta) = 0$. Then we see that there exists a solution $t(\sigma, \beta) = (q_1(z; \sigma, \beta), v_1(z; \sigma, \beta))$ of $S(t(\sigma, \beta), \sigma, \beta) = 0$ satisfying that $\lim_{\sigma \to 0} \|t(\sigma, \beta)\|_{C^2_{\sigma,0}} = 0$ uniformly in $\beta \in I_\delta$ and $t(\sigma, \beta)$ is uniformly continuous in σ and β relative to the norm of $C^2_{\sigma,0}$. Thus we obtain a solution $(\tilde{q}_0(z; \sigma, \beta) + q_1(z; \sigma, \beta), \tilde{v}_0(z; \sigma, \beta) + v_1(z; \sigma, \beta))$ with $z = x/\beta$ and $\sigma = \varepsilon/\beta$.

THEOREM 3.2. *There exist small positive constants ε_0 and δ such that for any $\varepsilon \in (0, \varepsilon_0)$, $\alpha \in (1-\delta, 1)$, and $\beta \in J_\delta$, (3.3) has a solution $(\tilde{q}(x; \varepsilon, \beta), \tilde{v}(x; \varepsilon, \beta))$ satisfying $\lim_{\varepsilon \to 0} \|\tilde{q}(x; \varepsilon, \beta) - 1\|_{C^2(I_\kappa)} = 0$ with $I_\kappa = [0, \beta - \kappa]$ for any small $\kappa > 0$ and $\lim_{\varepsilon \to 0} \varepsilon(d/dx)\tilde{q}(\beta; \varepsilon, \beta) = (d/d\eta)q_0(0; \beta)$ uniformly in β. Moreover, $(\tilde{q}, \tilde{v})$ is uniformly continuous in ε and β relative to the norm of $C^2_{\sigma,0}$ and $(d/d\eta)q_0(0; \beta)$ is strictly monotone increasing with respect to β.*

3.3. Patching of solutions. We proceed to construct a solution of (3.1), (3.2) by patching two solutions $(\tilde{q}, \tilde{v})$ and $(\hat{q}, \hat{v})$ as

$$(q(x; \varepsilon, \beta, \gamma), v(x; \varepsilon, \beta, \gamma)) = \begin{cases} (\tilde{q}(x; \varepsilon, \beta), \tilde{v}(x; \varepsilon, \beta)), & x \in I_\beta, \\ (\hat{q}(\xi; \varepsilon, \gamma), \hat{v}(\xi; \varepsilon, \gamma)), & \xi = (x-\beta)/\varepsilon \in R_\beta. \end{cases}$$

To do this, it is sufficient to show that at $x=\beta$, v is continuous and q is C^1-continuous. We first choose $\gamma=\gamma_\varepsilon(\beta)\equiv\tilde{v}(\beta;\varepsilon,\beta)=\beta+O(\varepsilon)$ to assure the continuity of v. Next consider $\Phi(\varepsilon,\beta)\equiv\varepsilon(d/dx)\tilde{q}(\beta;\varepsilon,\beta)-\varepsilon(d/dx)\hat{q}(0;\varepsilon,\gamma_\varepsilon(\beta))$. It follows immediately from Theorems 3.1 and 3.2 that $\Phi(\varepsilon,\beta)$ is uniformly continuous in $\varepsilon\in(0,\varepsilon_0)$ and $\beta\in I_\delta$ and

$$\Phi(0,\beta)\equiv\lim_{\varepsilon\to 0}\Phi(\varepsilon,\beta)=\frac{d}{d\eta}q_0(0;\beta)-\frac{d}{d\xi}\hat{q}_0(0,\beta).$$

Note (3.13), (3.14) and choose $\beta=\omega_0$ $(=c^*(a)/2\lambda)$. Then the unique existence of solutions (2.4) proves $\Phi(0,\omega_0)=0$. $\Phi(0,\beta)$ is strictly monotone increasing with respect to β since $(d/d\eta)q_0(0,\beta)$ and $(d/d\xi)\hat{q}_0(0,\beta)$ are strictly monotone increasing and decreasing, respectively. Therefore, it is easy to see that there exists a function $\beta=\beta(\varepsilon)$ defined for $\varepsilon\in(0,\varepsilon_0)$ for sufficiently small $\varepsilon_0>0$ that satisfies $\Phi(\varepsilon,\beta(\varepsilon))\equiv 0$ and $\lim_{\varepsilon\to 0}\beta(\varepsilon)=\omega_0$. For this $\beta(\varepsilon)$, $(q(x;\varepsilon,\beta(\varepsilon),\gamma_\varepsilon(\beta(\varepsilon))),\ v(x;\varepsilon,\beta(\varepsilon),\gamma_\varepsilon(\beta(\varepsilon))))$ is a solution of (3.1), (3.2) satisfying $\operatorname{supp}[q]=[0,\beta(\varepsilon)+\varepsilon\omega_\varepsilon(\gamma_\varepsilon(\beta(\varepsilon)))]\to[0,\omega_0]$ as $\varepsilon\to 0$. Thus, we have obtained a large solution $U(x,\varepsilon)=q(x;\varepsilon,\beta(\varepsilon),\gamma_\varepsilon(\beta(\varepsilon)))^\mu$ of (2.1), (2.2) satisfying all properties required in the theorem of § 2. This completes the first half of our theorem.

4. Existence of a small single cluster solution. For a small solution, we assume that $r\geqq 2\omega=2\varepsilon\sigma$ and set $\tilde{v}(\eta)=\int_0^\eta u(\eta')\,d\eta'$. Then, as in § 3, our problem (2.5), (2.6) is to find solutions (u,v) and σ of

$$\begin{aligned}&(u^m)_{\eta\eta}+2\varepsilon\lambda(\tilde{v}u)_\eta+f(u)=0,\\&\tilde{v}_\eta-u=0,\end{aligned}\qquad \eta\in J_\sigma=[0,\sigma),\tag{4.1}$$

with

$$\begin{aligned}&u_\eta(0)=0,\quad u(\sigma)=(u^m)_\eta(\sigma)=0,\quad u(\eta)>0\quad\text{for }\eta\in J_\sigma,\\&v(0)=0.\end{aligned}\tag{4.2}$$

In a slightly different manner from that of § 3, we introduce the new dependent variable (U,V,W) and the independent variable τ by

$$\begin{aligned}&U=u^{m-1},\quad V=\tilde{v},\quad W=(u^{m-1})_\eta+2\varepsilon\nu\tilde{v}=U_\eta+2\varepsilon\nu V,\\&\frac{d\eta}{d\tau}=U\quad\text{with }\tau|_{\eta=0}=0,\end{aligned}\tag{4.3}$$

where $\nu=\lambda(m-1)/m$. Now we rewrite (4.1), (4.2) as the first-order system

$$\begin{aligned}&U_\tau=U(W-2\varepsilon\nu V),\\&V_\tau=U^{1+\mu},\qquad\qquad\tau\in R_+=(0,\infty),\\&W_\tau=-\mu W(W-2\varepsilon\nu V)-F_0(U),\end{aligned}\tag{4.4}$$

with the conditions

$$\begin{aligned}&V(0)=W(0)=0,\qquad U(+\infty)=W(+\infty)=0,\\&W(\tau)<0\quad\text{for all }\tau>0,\end{aligned}\tag{4.5}$$

where $\mu=1/(m-1)$ and $F_0(U)=1/(1+\mu)U^{1-\mu}f(U^\mu)$. We should note that $(4.4)_\varepsilon$ has the one-dimensional critical manifold $\Omega_c\equiv\{(U,V,W)\,|\,U=0,\ W=0\}$ in the half-space $\{(U,V,W)\,|\,W\leqq 0\}$. In the following, we use vector notation for convenience. Let us write $(4.4)_\varepsilon$ as $y_\tau=f_\varepsilon(y)$ with $y=(U,V,W)$ and let $y_\varepsilon(\tau;\xi)$ denote the solution

of $(4.4)_\varepsilon$ satisfying $y_\varepsilon(0;\xi)=\xi$. We simplify this further by writing $\xi\cdot\tau=y_\varepsilon(\tau;\xi)$ and $\xi\cdot S=\{y\,|\,y=\xi\cdot\tau,\ \tau\in S\}$.

We first discuss the problem for $(4.4)_0$, namely,

$$(4.4)_0\qquad \begin{aligned} U_\tau &= UW,\\ V_\tau &= U^{1+\mu}, \qquad\qquad \tau\in R_+,\\ W_\tau &= -\mu W^2 - F_0(U), \end{aligned}$$

with (4.5). This is easily solved by phase plane analysis. Let γ_0 be a unique positive number satisfying

$$\int_0^{\gamma_0} U^\mu f(U^\mu)\,dU=\frac{1}{\mu}\int_0^{(\gamma_0)^\mu} u^{m-1}f(u)\,du=0$$

for each $a\in(0,a^*)$. Let $y_0(\tau;\gamma)=(U_0(\tau;\gamma),V_0(\tau;\gamma),W_0(\tau;\gamma))$ be a solution of $(4.4)_0$ satisfying $y_0(0;\gamma)=\xi_0$ with $\xi_0=(\gamma,0,0)$. Also, set $D=\{(U,V,W)\,|\,0<U<1,\ V>0,\ W<0\}$ and $E_0^-=\{(U,V,W)\,|\,0<U<a_0,\ V\geqq 0,\ W=0\}$. Then we have Lemma 4.1.

LEMMA 4.1. *For each $a\in(0,a^*)$, $y_0(\tau;\gamma_0)$ is a unique solution of $(4.4)_0$, (4.5) such that $U_0(\tau,\gamma_0)=O(\tau^{-2})$ and $W_0(\tau,\gamma_0)=O(\tau^{-1})$ as $\tau\to+\infty$, $\lim_{\tau\to\infty}V_0(\tau;\gamma_0)=V_0^*<+\infty$. Moreover, if $\gamma_0<\gamma<1$, $\lim_{\tau\to\infty}y_0(\tau;\gamma)=(0,V_0^\infty,-\infty)$ holds with some $V_0^\infty<+\infty$, while if $a_0<\gamma<\gamma_0$, there exists a unique finite $\tau(\gamma)>0$ such that $y_0(\tau;\gamma)\in D$ for $\tau\in(0,\tau(\gamma))$ and $y_0(\tau(\gamma);\gamma)\in E_0^-$.*

Proof. We can prove this lemma by the standard technique. In fact, the first and third equations of $(4.4)_0$ give the relation

$$\frac{1}{2}\frac{d}{d\tau}(U^\mu W)^2=\mu U^{2\mu-1}W^2U_\tau+U^{2\mu}WW_\tau=-U^{2\mu-1}F_0(U)U_\tau. \tag{4.6}$$

When we assume that there exists a monotone solution U, we can integrate this over $(0,\tau)$ as

$$\frac{1}{2}(U^\mu W)^2+\frac{1}{1+\mu}\int_\gamma^U S^\mu f(S^\mu)\,ds=0, \tag{4.7}$$

since $U(0)=\gamma$, $W(0)=0$. The condition $U(+\infty)=W(+\infty)=0$ requires $\gamma=\gamma_0$, so we have

$$\frac{dU}{d\tau}=-U^{1-\mu}\left[\frac{2}{1+\mu}\int_U^{\gamma_0}s^\mu f(s^\mu)\,ds\right]^{1/2}\equiv-\Phi(U;\gamma_0). \tag{4.8}$$

Integrating (4.8) with $U(0)=\gamma_0$, we see that $U(\tau)$ satisfies $\tau=\int_U^{\gamma_0}\Phi(s;\gamma_0)^{-1}\,ds$. There exists a unique function $U_0(\tau;\gamma_0)$ satisfying this relationship. Conversely, such a function gives a solution of $(4.4)_0$, (4.5). To see the decay rate of U_0, it suffices to examine (4.8) near $U=0$ as

$$\begin{aligned}\frac{dU}{d\tau}&=-U^{1-\mu}\left[-\frac{2}{1+\mu}\int_0^U s^\mu f(s^\mu)\,ds\right]^{1/2}\\ &=-\sqrt{\frac{a}{\mu(1+\mu)}}\,U^{3/2}(1+O(U^\mu)).\end{aligned} \tag{4.9}$$

Integrating this proves the estimate $U_0(\tau;\gamma_0)=O(\tau^{-2})$ as $\tau\to+\infty$, which also assures $W_0(\tau;\gamma_0)=O(\tau^{-1})$ as $\tau\to+\infty$.

The behavior of solutions $y_0(\tau;\gamma)$ is easily derived by examining relation (4.7) in a way similar to the above. This completes the proof.

Remark 4.1. We see $U_0 \in L^1(R_+)$ since U_0 satisfies $0 < U_0 \leqq \gamma$ for $\tau \in R_+$ and $U_0 = O(\tau^{-2})$ as $\tau \to +\infty$. Hence, $\eta = \phi_0(\tau) \equiv \int_0^\tau U_0(\tau'; \gamma_0)\, d\tau'$ is monotone increasing and $\lim_{\tau\to\infty} \eta(\tau) = \sigma_0 < +\infty$, so that the inverse function $\tau = \phi_0^{-1}(\eta)$ is defined on $[0, \sigma_0)$ and satisfies $\phi_0^{-1}(0) = 0$ and $\lim_{\eta\to\sigma_0} \phi_0^{-1}(\eta) = +\infty$. Set $u_i(\eta; \sigma_0) = U_0(\phi_0^{-1}(\eta); \gamma_0)^\mu$. Then it is a unique solution of (2.7), (2.6) since

$$(u_i^m)_\eta(\eta)\big|_{\eta=\sigma_0} = (U_0^{1+\mu})_\tau \frac{d\tau}{d\eta}\bigg|_{\tau=+\infty} = (1+\mu) U_0^\mu W_0|_{\tau=+\infty} = 0.$$

Next we will prove the existence of solutions of $(4.4)_\varepsilon$, (4.5) for $\varepsilon > 0$. To do this, we apply the variant of the Wazewski theorem formulated by Dunbar in [6]. We begin by constructing a Wazewski set $\Omega \subset R^3$. For small $\delta_1 > 0$, define $H(U)$ by

$$H(U) = \begin{cases} -\delta_1 U & (0 \leqq U \leqq a_0/2), \\ \delta_1(U - a_0) & (a_0/2 < U \leqq a_0), \\ 0 & (a_0 < U \leqq 1). \end{cases}$$

For a sufficiently small positive δ_2 (<1) and a large C, define the set Ω bounded by

$$\Omega_1^+ = \{(U, V, W) \mid 0 < U \leqq 1,\ V = 0,\ -\sqrt{U/\delta_2} < W < H(U)\},$$
$$\Omega_2^+ = \{(U, V, W) \mid a_0 < U \leqq 1,\ 0 \leqq \varepsilon V \leqq C + \varepsilon(1-U),\ W = 0\},$$
$$\Omega_3^+ = \{(U, V, W) \mid U = 1,\ 0 \leqq \varepsilon V \leqq C,\ -1/\sqrt{\delta_2} < W \leqq 0\},$$
$$\Omega_4^+ = \{(U, V, W) \mid 0 < U < 1,\ \varepsilon V = C + \varepsilon(1-U),\ -\sqrt{U/\delta_2} < W < H(U)\},$$
$$\Omega_1^- = \{(U, V, W) \mid 0 < U \leqq a_0,\ 0 \leqq \varepsilon V \leqq C + \varepsilon(1-U),\ W = H(U)\},$$
$$\Omega_2^- = \{(U, V, W) \mid U = \delta_2 W^2,\ 0 \leqq \varepsilon V \leqq C + \varepsilon(1-U),\ -1/\sqrt{\delta_2} \leqq W < 0\},$$
$$\Omega_0 = \{(U, V, W) \mid U = W = 0,\ 0 \leqq \varepsilon V \leqq C + \varepsilon\},$$

where δ_i $(i = 1, 2)$ and C will be specified later so that Ω becomes a Wazewski set (see Fig. 5). The boundary $\partial\Omega$ of Ω is represented by $\partial\Omega = \Omega_0 \cup \Omega^+ \cup \Omega_1^- \cup \Omega_2^-$, where $\Omega^+ = \cup_{i=1}^4 \Omega_i^+$. Note that Ω_1^- and Ω_2^- are disjoint. Let Ω^- be the immediate exit set of Ω, that is, for all $y_0 \in \Omega^-$, $y_0 \cdot [0, \tau) \not\subset \Omega$ for any $\tau > 0$. Then, we have the following lemma.

LEMMA 4.2. *Let δ_i $(i = 1, 2)$ be sufficiently small, and let C be sufficiently large. Then, $\Omega^- = \Omega_1^- \cup \Omega_2^-$.*

Proof. Set $(\Omega^-)^c = \partial\Omega \backslash \Omega^-$. It is obvious that $\Omega_i^+ \subset (\Omega^-)^c$ for $i = 1, 2, 3$. We have that $\Omega_0 \subset (\Omega^-)^c$ since any point in Ω_0 is a critical point of $(4.4)_\varepsilon$. Therefore, we may only consider Ω_4^+, Ω_1^-, and Ω_2^-. First consider Ω_4^+. $n_4 = (1, 1, 0)$ is an outward normal for Ω at $y \in \Omega_4^+$. For $y \in \Omega_4^+$, $0 < U < 1$, $W < 0$, and $\varepsilon V = C + \varepsilon(1-U) > C$, so that the inner product $n_4 \cdot f_\varepsilon(y)$ is estimated as

$$n_4 \cdot f_\varepsilon(y) = U(W - 2\varepsilon\nu V) + U^{1+\mu} < (1 - 2\nu C) U < 0,$$

if we choose $C > 1/(2\nu)$. This implies $\Omega_4^+ \subset (\Omega^-)^c$ for any $C > 1/(2\nu)$.

The outward normal at $y \in \Omega_2^-$ is given by $n_2 = (-1, 0, 2\delta_2 W)$. For $y \in \Omega_2^-$, $V \geqq 0$, $W < 0$, $0 < U \leqq 1$, and $U = \delta_2 W^2$. Hence, we have

$$\begin{aligned} n_2 \cdot f_\varepsilon(y) &= -U(W - 2\varepsilon\nu V) - 2\delta_2 W^2(W - 2\varepsilon\nu V) - 2\delta_2 W F_0(U) \\ &= -(1+2\mu)\delta_2 W^2(W - 2\varepsilon\nu V) - 2\delta_2 W F_0(\delta_2 W^2) \\ &> -(1+2\mu)\delta_2 W^3 + 2\delta_2^2 W^3 k_0 \\ &= \delta_2 W^3[2\delta_2 k_0 - (1+2\mu)], \end{aligned}$$

where $k_0 = -\inf_{0<U\leqq 1}(F_0(U)/U)$ $(= a/(1+\mu))$. If we choose $\delta_2 < (1+2\mu)/2k_0$, $n_2 \cdot f_\varepsilon(y) > 0$, which proves $\Omega_2^- \subset \Omega^-$. Consider Ω_1^- and define $E_\pm$ by $E_+ = \{(U, V, W) \in \Omega_1^- \mid 0 < U < a_0/2\}$ and $E_- = \{(U, V, W) \in \Omega_1^- \mid a_0/2 \leqq U \leqq a_0\}$. The outward

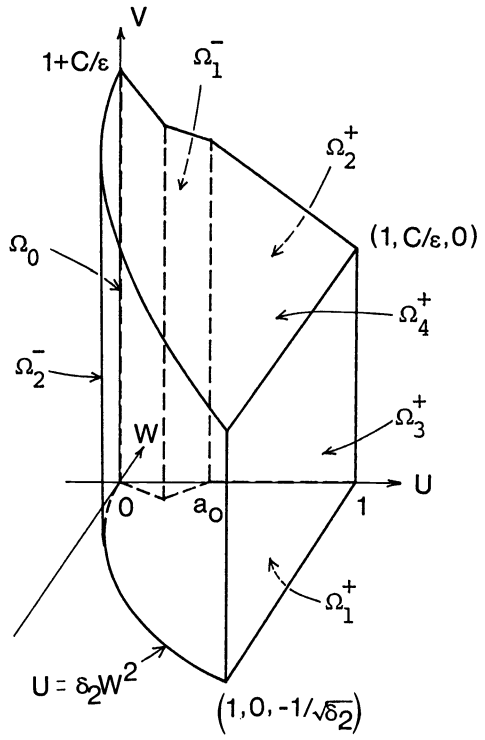


FIG. 5. *The Wazewski set for a small cluster solution.*

normals for Ω at $y \in E_\pm$ are $n_\pm = (\pm\delta_1, 0, 1)$, respectively. For $y \in E_+$, $W = -\delta_1 U$, $0 < U < a_0/2$, and $\varepsilon V < C + \varepsilon < C + 1$, so that

$$\begin{aligned} n_+ \cdot f_\varepsilon(y) &= \delta_1 U(W - 2\varepsilon\nu V) - \mu W(W - 2\varepsilon\nu V) - F_0(U) \\ &> -\delta_1(1+\mu)U\{\delta_1 U + 2V(C+1)\} + k_1 U > 0 \end{aligned}$$

with $k_1 = \inf_{0<U<a_0/2}(-F_0(U)/U)$, if we choose $\delta_1 < k_1/(1+\mu)\{1+2\nu(C+1)\}$. Analogously, for $y \in E_-$ we have

$$\begin{aligned} n_- \cdot f_\varepsilon(y) &= -\delta_1 U(W - 2\varepsilon\nu V) - \mu W(W - 2\varepsilon\nu V) - F_0(U) \\ &> -\mu\delta_1(a_0 - U)\{\delta_1(a_0 - U) + 2\nu(C+1)\} + k_2(a_0 - U) > 0, \end{aligned}$$

with $k_2 = \inf_{a_0/2<U<a_0}(-F_0(U)/(a_0 - U))$, if we choose $\delta_1 < k_2/\mu\{1+2\nu(C+1)\}$. Hence, we have $\Omega_1^- = E_+ \cup E_- \subset \Omega^-$. This completes the proof.

Let δ_0 be a small positive constant satisfying $\Sigma_U \equiv [\gamma_0 - \delta_0, \gamma_0 + \delta_0] \subset (a_0, 1)$, and set $\Sigma \equiv \{(U, V, W) \mid U \in \Sigma_U,\ V = W = 0\}$. Let $\Sigma^0 = \{\xi_0 \in \Sigma \mid$ with a $\tau_0 = \tau_0(\xi_0)$ such that $\xi_0 \cdot \tau_0 \in \Omega\}$. For $\xi_0 \in \Sigma^0$, define $T(\xi_0) = \sup\{\tau \mid \xi_0 \cdot [0, \tau] \subseteq \Omega\}$. Suppose $\xi_0 \in \Sigma$ and $\xi_0 \cdot [0, \tau] \subseteq \text{cl}(\Omega)$, where cl (Ω) denotes the closure of Ω. Then $\xi_0 \cdot [0, \tau] \subset \Omega$ since Ω is closed. Suppose $\xi_0 \in \Sigma$, $\tau < T(\xi_0)$, $y_\varepsilon(\tau; \xi_0) \notin \Omega^-$. Then, $y_\varepsilon(\tau; \xi_0) \in \text{int}\,\Omega \cup \Omega_0 \cup \Omega^+$. Here, Ω_0 consists only of critical points, so $y_\varepsilon(\tau; \xi_0) \notin \Omega_0$. After some calculation, we can show that for any $\xi \in \Omega^+$, there exists $\tau_0 > 0$ such that $\xi \cdot \tau' \in \Omega$ for all $\tau' \in [-\tau_0, 0)$, which proves $y_\varepsilon(\tau; \xi_0) \notin \Omega^+$. Hence, we have $y_\varepsilon(\tau; \xi_0) \in \text{int}\,\Omega$. This implies that there is an open set S_τ about $\xi_0 \cdot \tau$ disjoint from Ω^-. Thus, it turns out that Ω is a Wazewski set.

Next, consider the mapping $\mathcal{T}(\xi_0) = \xi_0 \cdot T(\xi_0)$ from Σ^0 into Ω^-. For given γ_1, γ_2 satisfying $a_0 < \gamma_1 < \gamma_0 < \gamma_2 < 1$, set $\xi_i = (\gamma_i, 0, 0)$ $(i = 1, 2)$. Then from Lemma 4.1 and

the continuous dependence of a parameter ε, it follows that there exists a small $\varepsilon_0 > 0$ such that for each $\varepsilon \in [0, \varepsilon_0)$, $y_\varepsilon(\tau; \xi_1)$ and $y_\varepsilon(\tau; \xi_2)$ intersect Ω_1^- and Ω_2^- in finite time, respectively, that is, $\mathcal{T}(\xi_1) \in \Omega_1^-$ and $\mathcal{T}(\xi_2) \in \Omega_2^-$. Note that Σ is compact and intersects a trajectory of $(4.4)_\varepsilon$ only once in Ω. Then, if $\Sigma = \Sigma^0$, Proposition 1 of [6] proves that $\mathcal{T}$ is a homeomorphism of the connected set Σ to its image in the disconnected set Ω^-. This is a contradiction, so that $\Sigma \neq \Sigma^0$. Therefore, there is a $\xi_\varepsilon = (\gamma_\varepsilon, 0, 0) \in \Sigma$ such that the solution $y_\varepsilon(\tau; \xi_0)$ of $(4.4)_\varepsilon$ remains in Ω for all $\tau \geqq 0$.

LEMMA 4.3. *There is a small $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$, $(4.4)_\varepsilon$, (4.5) has a solution $y_\varepsilon(\tau; \xi_\varepsilon) = (U_\varepsilon(\tau; \gamma_\varepsilon), V_\varepsilon(\tau; \gamma_\varepsilon), W_\varepsilon(\tau; \gamma_\varepsilon))$ satisfying*

$$\lim_{\tau \to \infty} y_\varepsilon(\tau; \xi_\varepsilon) = (0, V^*, 0), \tag{4.10}$$

where V^* satisfies $0 < C_1 < V^* < C_2$. Moreover, $v_\varepsilon(\tau; \xi_\varepsilon)$ satisfies

$$\lim_{\tau \to \infty} U_\varepsilon(\tau; \gamma_\varepsilon) / W_\varepsilon(\tau; \gamma_\varepsilon) = \frac{2(1+\mu)\nu V^*}{F_0'(0)} \varepsilon. \tag{4.11}$$

Proof. It suffices for us to show (4.10) and (4.11). Any point $y \in \Omega \backslash \Omega_0$ is an ordinary point of $(4.4)_\varepsilon$, where $U_\tau < 0$ and $V_\tau > 0$, so that $y_\varepsilon(\tau; \xi_\varepsilon)$ must approach some point $(0, V^*, 0) \in \Omega_0$ as $\tau \to \infty$. This implies that for any small ε there exists some $\tau_\varepsilon > 0$ satisfying $U_\varepsilon(\tau_\varepsilon) = a_0/2$ and $0 < \tau_1 < \tau_\varepsilon < \tau_2$ with some constants τ_1, τ_2 independent of ε. Since $W_\varepsilon < -\delta_1 U_\varepsilon$ for $\tau \geqq \tau_\varepsilon$, it follows from the first equation $(4.4)_\varepsilon$ that

$$\frac{dU_\varepsilon}{d\tau} = U_\varepsilon(W_\varepsilon - 2\varepsilon\nu V_\varepsilon) \leqq U_\varepsilon W_\varepsilon \leqq -\delta_1 U_\varepsilon^2.$$

Integrating this over (τ_ε, τ), we have

$$U_\varepsilon(\tau) \leqq \frac{a_0}{2 + a_0\delta_1(\tau - \tau_\varepsilon)}, \tag{4.12}$$

where $U_\varepsilon(\tau)$ simply denotes $U_\varepsilon(\tau; \gamma_\varepsilon)$. Then the second equation of $(4.4)_\varepsilon$ shows

$$\begin{aligned} V_\varepsilon(\tau) &= \int_0^{\tau_\varepsilon} U_\varepsilon(\tau')^{1+\mu}\, d\tau' + \int_{\tau_\varepsilon}^{\tau} U_\varepsilon(\tau')^{1+\mu}\, d\tau' \\ &\leqq \tau_\varepsilon + \frac{a_0^\mu}{\delta_1 \mu} \frac{1}{[2 + a_0\delta_1(\tau - \tau_\varepsilon)]^\mu} \leqq C_2 \end{aligned}$$

with some positive constant C_1 independent of ε. It is obvious that there is $C_1 > 0$ independent of ε such that $V_\varepsilon(\tau_\varepsilon) \geqq C_1$. Since $V_\varepsilon(\tau)$ is monotone increasing, $V_\varepsilon(\tau)$ has a unique limit V^* satisfying $C_1 \leqq V^* \leqq C_2$ as $\tau \to \infty$.

Let us prove (4.11). We first note that for any $\delta > 0$, there exists a large τ_0 such that $V^* - \delta \leqq V_\varepsilon(\tau) < V^*$ for all $\tau \geqq \tau_0$. The first and third equations of $(4.4)_\varepsilon$ are written as follows:

$$\begin{aligned} U_\tau &= -2\varepsilon\nu V^* U + g_1, \\ W_\tau &= 2\varepsilon\mu\nu V^* W - F_0'(0) U + g_2, \end{aligned}$$

where $g_1 = UW - 2\varepsilon\nu(V - V^*)U$ and $g_2 = -\mu W^2 + 2\varepsilon\nu\mu(V - V^*)W - \{F_0(U) - F_0'(0)U\}$. Set $Z = {}^t(U, W)$ and $G(Z, V) = {}^t(g_1, g_2)$. By the linear transformation $Z = P\tilde{Z}$ with

$$P = \begin{pmatrix} 1 & 0 \\ F_0'(0)/2\varepsilon\nu V^*(1+\mu) & 1 \end{pmatrix},$$

we have the equation for $\tilde{Z}$:

$$\frac{d\tilde{Z}}{d\tau} = \begin{pmatrix} -2\varepsilon\nu V^* & 0 \\ 0 & 2\varepsilon\mu\nu V^* \end{pmatrix} \tilde{Z} + \tilde{G}(\tilde{Z}, \tau),$$

where $\tilde{G}(\tilde{Z}, \tau) = P^{-1}G(P\tilde{Z}, V(\tau))$. By applying the result in Theorems 4.1 and 4.2 of [5] to this equation, we can prove (4.11), which completes the proof.

Now we can show the properties of a small solution stated in Theorem 2.1. To do so, we need further information on the solution. Consider the plane $E_1 = \{(U, V, W) \mid 0 < U \leqq \varepsilon^2,\ 0 \leqq V \leqq C_2,\ W = -\delta_2 U/\varepsilon$ having the normal vector $m_1 = (\delta_2, 0, \varepsilon)$ with $\delta_2 \in (0, 1)$. Then the inner product $m_1 \cdot f_\varepsilon$ is estimated as

$$\begin{aligned} m_1 \cdot f_\varepsilon &= \delta_2 U(W - 2\varepsilon\nu V) - \varepsilon\mu W(W - 2\varepsilon\nu V) - \varepsilon F_0(U) \\ &= U\{\delta_2(1+\mu)(W - 2\varepsilon\nu V) + \varepsilon(-F_0(U)/U)\} \\ &> U_\varepsilon\{k_1 - \delta_2(1+\mu)(\delta_2 + 2\varepsilon\nu C_2)\} > 0, \end{aligned}$$

if we choose $\delta_2 < k_1/(1+\mu)(1+2\nu C_2)$. Also, consider the surface $E_2 = \{(U, V, W) \mid \varepsilon_2 < U \leqq a_0/2,\ 0 \leqq V \leqq C_2,\ W^2 = \delta_2^2 U,\ W < 0\}$ having the normal vector $m_2 = (\delta_2^2, 0, -2W)$. Then for E_2,

$$\begin{aligned} m_2 \cdot f_\varepsilon &= \delta_2^2 U(W - 2\varepsilon\nu V) + 2\mu W^2(W - 2\varepsilon\nu V) + 2WF_0(U) \\ &\geqq U[(1+2\mu)\delta_2^2(W - 2\varepsilon\nu C_2) - 2Wk_1] \\ &\geqq U[\{(1+2\mu)\delta_2^2 - k_1\}W - \delta_2\varepsilon\{2(1+2\mu)\nu C_2\delta_2 - k_1\}] \\ &> 0 \end{aligned}$$

for sufficiently small positive δ_2. Here we have used the estimate $V \leqq C_2$. Together with (4.11), these two estimates prove that the orbit $y_\varepsilon(\tau; \xi_\varepsilon)$ cannot intersect $E_1 \cup E_2$. That is, we choose U as the independent variable in place of τ through $U = U_\varepsilon(\tau; \gamma_\varepsilon) \equiv \psi_\varepsilon(\tau)$ and denote the solution orbit as S_ε: $(U, V_\varepsilon(U), W_\varepsilon(U)) = (U, V_\varepsilon(\psi_\varepsilon^{-1}(U); \gamma_\varepsilon), W_\varepsilon(\psi_\varepsilon^{-1}(U); \gamma_\varepsilon))$ defined for $U \in (0, \gamma_\varepsilon)$, where ψ_ε^{-1} is the inverse function of ψ_ε. Then $W_\varepsilon(U) < -\delta_2 U/\varepsilon$ for $0 < U \leqq \varepsilon^2$ and $W_\varepsilon(U) < -\delta_2\sqrt{U}$ for $\varepsilon^2 \leqq U \leqq a_0/2$. Also, we denote the solution orbit of $y_0(\tau; \gamma_0)$ as S_0: $(U, V_0(U), W_0(U)) = (U, V_0(\psi_0^{-1}(U); \gamma_0), W_0(\psi_0^{-1}(U); \gamma_0))$ for $U \in (0, \gamma_0)$, where $U = U_0(\tau, \gamma_0) \equiv \psi_0(\tau)$ and its inverse is ψ_0^{-1}.

Let us study the ε-dependence of S_ε. It follows from $(4.4)_\varepsilon$ that for $U \in (0, \underline{\gamma})$ with $\underline{\gamma} = \min(\gamma_0, \gamma_\varepsilon)$,

$$\frac{1}{2}(U^\mu W_\varepsilon)^2 - \frac{1}{2}(U^\mu W_0)^2 = -\frac{2\varepsilon\nu}{1+\mu}\int_0^U s^{\mu+1} f(s^\mu) V_\varepsilon(s)\left(\frac{dU_\varepsilon}{d\tau}(s)\right)^{-1} ds \tag{4.13}$$

(see (4.6) and (4.7)). Note that for small $\delta > 0$ there is a large τ_0 such that $V_\varepsilon \geqq V^* - \delta \equiv V_\delta$ for $\tau \geqq \tau_0$ and $0 < \inf_{0<\varepsilon<\varepsilon_0} U_\varepsilon(\tau_0) \equiv U_1 < a_0/2$ with small $\varepsilon_0 > 0$. Then $|dU_\varepsilon/d\tau(U)| = |U(W_\varepsilon - 2\varepsilon\nu V_\varepsilon)| \geqq U(\delta_2 U/\varepsilon + 2\varepsilon\nu V_\delta)$ for $0 < U \leqq \varepsilon^2$ and $|dU_\varepsilon/d\tau(U)| \geqq \delta_2 U^{3/2}$ for $\varepsilon^2 \leqq U \leqq U_1$, so that for $0 < U \leqq \varepsilon^2$,

$$\begin{aligned} I_f &\equiv \left| \int_0^U s^{\mu+1} f(s^\mu) V_\varepsilon\left(\frac{dU_\varepsilon}{d\tau}\right)^{-1} d\tau \right| \leqq \frac{V^* K_1}{\delta_2}\int_0^U \frac{\varepsilon s^{2\mu+1}\, ds}{s(s+\varepsilon^2 a)} \\ &\leqq \frac{V^* K_1}{\delta_2} U^{2\mu}\varepsilon \log\left(1 + \frac{U}{\varepsilon^2 a}\right) \leqq C_3\varepsilon U^{2\mu}, \end{aligned}$$

with $K_1 = \sup_{0<u<1} |f'(U)|$ and $a = 2\nu V_\delta/\delta_2$. For $\varepsilon^2 < U \leqq U_1$, we know

$$\begin{aligned} I_f &\leqq C_3\varepsilon(\varepsilon^2)^{2\mu} + V^* K_1 \int_{\varepsilon^2}^U \frac{s^{2\mu}\, ds}{\delta_2\sqrt{s} + 2\varepsilon\nu V_\delta} \\ &\leqq \left(C_3\varepsilon + \frac{V^* K_1}{2\delta_2}\sqrt{U}\right) U^{2\mu} \leqq C_4 U^{2\mu}. \end{aligned}$$

Furthermore, we easily see that for small $\delta'>0$, $|dU_\varepsilon/d\tau(U)^{-1}|\leqq C_5$ for $U\in[U_1, \gamma_\varepsilon-\delta']$ and $|dU_\varepsilon/d\tau(U)^{-1}|\leqq C_6/\sqrt{\gamma_\varepsilon-U}$ for $U\in[\gamma_\varepsilon-\delta', \gamma_\varepsilon)$. Applying all these estimates to I_f, we have $|W_\varepsilon(U)^2-W_0(U)^2|\leqq C_7\varepsilon$ for all $U\in(0, \underline{\gamma})$, so that

$$\sup_{0<U<\underline{\gamma}}|W_\varepsilon(U)-W_0(U)|\leqq C_8\sqrt{\varepsilon}. \tag{4.14}$$

At the same time, this proves $\lim_{\varepsilon\to 0}\gamma_\varepsilon=\gamma_0$. Since $U_\varepsilon(0;\gamma_\varepsilon)=\gamma_\varepsilon\to U_0(0;\gamma_0)=\gamma_0$ as $\varepsilon\to 0$ and $V_\varepsilon(0;\gamma_\varepsilon)=V_0(0;\gamma_0)=W_\varepsilon(0;\gamma_\varepsilon)=W_0(0;\gamma_0)=0$, the continuous dependence of solutions on initial values and parameters shows that $(U_\varepsilon(\tau;\gamma_\varepsilon), V_\varepsilon(\tau;\gamma_\varepsilon), W_\varepsilon(\tau;\gamma_\varepsilon))\to(U_0(\tau;\gamma_0), V_0(\tau;\gamma_0), W_0(\tau;\gamma_0))$ uniformly on any compact subset of R_+ as $\varepsilon\to 0$. Also note that (4.12) and Lemma 4.1 show $|U_\varepsilon(\tau;\gamma_\varepsilon)|, |U_0(\tau;\gamma_0)|\leqq C_9/\tau$ as $\tau\to\infty$. Therefore, we have

$$\lim_{\varepsilon\to 0}\sup_{\tau\in R_+}|U_\varepsilon(\tau;\gamma_\varepsilon)-U_0(\tau;\gamma_0)|=0. \tag{4.15}$$

To prove the compactness of support, let us consider the difference between $\eta_\varepsilon=\phi_\varepsilon(\tau)=\int_0^\tau U_\varepsilon(\tau';\gamma_\varepsilon)\,d\tau'$ and $\eta=\phi_0(\tau)=\int_0^\tau U_0(\tau';\gamma_0)\,d\tau'$, which is written as follows:

$$\begin{aligned}\phi_\varepsilon(\tau)-\phi_0(\tau)&=\int_0^{\tau_0}[U_\varepsilon(\tau';\gamma_\varepsilon)-U_0(\tau';\gamma_0)]\,d\tau'+\int_{U_\varepsilon(\tau_0)}^{U_0(\tau_0)}\frac{dU}{W_\varepsilon(U)-2\varepsilon\nu V_\varepsilon(U)}\\&\quad+\int_{U_0(\tau_0)}^{U_\varepsilon(\tau)}\left[\frac{1}{W_\varepsilon(U)-2\varepsilon\nu V_\varepsilon(U)}-\frac{1}{W_0(U)}\right]dU-\int_{U_\varepsilon(\tau)}^{U_0(\tau)}\frac{dU}{W_0(U)}\\&\equiv I_1+I_2+I_3+I_4,\end{aligned}$$

for sufficiently large τ_0. For any finite τ_0 fixed, I_1 and I_2 obviously tend to zero as $\varepsilon\to 0$. I_3 is evaluated as follows:

$$|I_3|\leqq\left(\sup_{0<U<U_0(\tau_0)}|W_\varepsilon(U)-W_0(U)|+2\varepsilon\nu V^*\right)\int_0^{U_0(\tau_0)}\frac{dU}{W_0(U)[W_\varepsilon(U)-2\varepsilon\nu V_\delta]}.$$

Since (4.9) assures us that $W_0\leqq-\delta_3\sqrt{U}$ for $U\in(0, U_0(\tau_0))$ with small $\delta_3>0$, elementary calculus shows that

$$\begin{aligned}\int_0^{U_0(\tau_0)}\frac{dU}{W_0(W_\varepsilon-2\varepsilon\nu V_\delta)}&\leqq\int_0^{\varepsilon_2}\frac{\varepsilon\,dU}{\delta_2\delta_3\sqrt{U}(U+\varepsilon^2 b)}+\int_{\varepsilon^2}^{U_0(\tau_0)}\frac{dU}{\delta_2\delta_3\sqrt{U}(\sqrt{U}+\varepsilon b)}\\&=\frac{2}{\delta_2\delta_3\sqrt{b}}\tan^{-1}\frac{1}{\sqrt{b}}+\frac{2}{\delta_2\delta_3}\log\frac{\sqrt{U_0(\tau_0)}+\varepsilon b}{(1+b)\varepsilon},\end{aligned}$$

where $b=2\nu V_\delta/\delta_2$. Together with (4.14), this proves $|I_3|\leqq C_{10}\sqrt{\varepsilon}\,|\log\varepsilon|\to 0$ as $\varepsilon\to 0$. It follows from (4.15) that

$$|I_4|\leqq\frac{1}{\delta_3}\left|\int_{U_\varepsilon(\tau)}^{U_0(\tau_0)}dU/\sqrt{U}\right|=|\sqrt{U_0(\tau)}-\sqrt{U_\varepsilon(\tau)}|/\delta_3\to 0$$

uniformly in $\tau\in R_+$ as $\varepsilon\to 0$. Finally, we have

$$\lim_{\varepsilon\to 0}\sup_{\tau\in R_+}|\phi_\varepsilon(\tau)-\phi_0(\tau)|=0, \tag{4.16}$$

which means, in particular, that $|\sigma_\varepsilon-\sigma_0|\equiv|\phi_\varepsilon(+\infty)-\phi_0(+\infty)|\to 0$ as $\varepsilon\to 0$.

Next, let us show (2.10). Let $\tau=\phi_\varepsilon^{-1}(\eta)$ be the inverse function of $\eta=\phi_\varepsilon(\tau)$ defined on $[0, \sigma_\varepsilon)$, and let $U_\varepsilon(\eta)$ be $U_\varepsilon(\phi_\varepsilon^{-1}(\eta);\gamma_\varepsilon)$. It is obvious that

$$u=\begin{cases}U_\varepsilon(\eta)^\mu & (0\leqq\eta<\sigma_\varepsilon),\\ 0 & (\eta\geqq\sigma_\varepsilon)\end{cases}$$

is a solution of (4.1), (4.2). Set $\varrho = \min(\sigma_\varepsilon, \sigma_0)$ and consider $U_\varepsilon(\eta) - U_0(\eta)$ for $\eta \in [0, \varrho)$, which is represented as

$$\begin{aligned} U_\varepsilon(\eta) - U_0(\eta) &= U_\varepsilon(\phi_\varepsilon^{-1}(\eta); \gamma_\varepsilon) - U_0(\phi_0^{-1}(\eta); \gamma_0) \\ &= U_\varepsilon(\phi_0^{-1}(\eta); \gamma_\varepsilon) - U_0(\phi_0^{-1}(\eta); \gamma_0) + U_\varepsilon(\phi_\varepsilon^{-1}(\eta); \gamma_\varepsilon) \\ &\quad - U_\varepsilon(\phi_0^{-1}(\eta); \gamma_\varepsilon). \end{aligned}$$

The first difference on the right-hand side tends to zero uniformly in $\tau = \phi_0^{-1}(\eta) \in R_+$ as $\varepsilon \to 0$, so that we only consider the second difference as follows:

$$\begin{aligned} \Delta &= U_\varepsilon(\phi_\varepsilon^{-1}(\eta); \gamma_\varepsilon) - U_\varepsilon(\phi_0^{-1}(\eta); \gamma_\varepsilon) = U(\phi_\varepsilon^{-1}(\phi_0(\tau)); \gamma_\varepsilon) - U_\varepsilon(\tau; \gamma_\varepsilon) \\ &= U_\varepsilon(\phi_\varepsilon^{-1}(\eta); \gamma_\varepsilon) - U_\varepsilon(\phi_\varepsilon^{-1}(\eta_\varepsilon); \gamma_\varepsilon) \end{aligned}$$

with $\eta_\varepsilon = \phi_\varepsilon(\tau)$. Hence, noting $d\phi_\varepsilon^{-1}/d\eta(\eta) = [(d\phi_\varepsilon/d\tau)(\tau_\varepsilon)]^{-1} = U_\varepsilon(\tau_\varepsilon; \gamma_\varepsilon)^{-1}$ with $\tau_\varepsilon = \phi_\varepsilon^{-1}(\eta)$, we easily have

$$\begin{aligned} |\Delta| &= \left| \frac{dU_\varepsilon}{d\tau}(\bar{\tau}_\varepsilon, \gamma_\varepsilon) \frac{d\phi_\varepsilon^{-1}}{d\eta}(\bar{\eta})(\eta - \eta_\varepsilon) \right| \leqq C_{11} |\phi_\varepsilon(\tau) - \phi_0(\tau)| \\ &\to 0 \quad \text{uniformly in } \tau \in R_+ \quad \text{as } \varepsilon \to 0, \end{aligned}$$

where $\bar{\eta} = \phi_\varepsilon(\bar{\tau}_\varepsilon)$ is an intermediate value between η and η_ε. Thus, we have $\lim_{\varepsilon \to 0} \sup_{0 < \eta < \varrho} |U_\varepsilon(\eta) - U_0(\eta)| = 0$. This completes the proof of Theorem 2.1.

5. Appendix.

5.1. Proof of Lemma 3.1. Let $(q_{0\zeta}, p_{0\zeta}) = (\varphi_1, \varphi_2)$. Differentiating $(3.6)_\varepsilon$ with respect to ζ, we see that $\varphi = {}^t(\varphi_1, \varphi_2)$ satisfies the linearized equations of $(3.6)_0$, that is, $Q_1 \varphi = 0$. Since $\varphi_1(\zeta) < 0$ for all $\zeta \geqq 0$, by reducing the order of the system, we have the linearly independent solution $\psi = {}^t(\psi_1, \psi_2)$ of $Q_1 \psi = 0$ as

$$\begin{aligned} &\psi_1(\zeta) = \varphi_1(\zeta)\chi(\zeta), \\ &\psi_2(\zeta) = \varphi_2(\zeta)\chi(\zeta) + \exp\left(- \int_0^\zeta A(\zeta')\, d\zeta' \right), \end{aligned} \tag{5.1}$$

where

$$A(\zeta) = \frac{\varphi_2(\zeta)}{\varphi_1(\zeta)} q_0(\zeta) + 2\mu \left(p_0(\zeta) + \frac{\lambda\gamma}{m\mu} \right)$$

and

$$\chi(\zeta) = \int_0^\zeta \left[\frac{q_0(\zeta')}{\varphi_1(\zeta')} \exp\left(- \int_0^{\zeta'} A(\zeta'')\, d\zeta'' \right) \right] d\zeta'.$$

As $\zeta \to +\infty$,

$$B(+\infty) = \begin{pmatrix} -2\lambda\gamma/m\mu & 0 \\ -F_0'(0) & 2\lambda\gamma/m \end{pmatrix},$$

so that its eigenvalues are $\rho_+ = 2\lambda\gamma/m$ and $\rho_- = -2\lambda\gamma/m\mu$, and the corresponding eigenvectors are ${}^t(0, 1)$ and ${}^t(2\lambda\gamma, F_0'(0))$, respectively. By the standard argument in the theory of ordinary differential equations, noting $F_0(q) \in C^{1,\mu}$, where $C^{1,\mu}$ is the set of functions whose first derivative is Hölder continuous with the exponent μ, we see that q_0 and $p_0 + (2\lambda\gamma/m\mu)$ belong to the class $X^2_{-\rho_-}(R_+)$, which means $\varphi_1, \varphi_2 \in X^1_{-\rho_-}(R_+)$. Knowing this, we have Proposition 5.1.

PROPOSITION 5.1. $|\psi_2(\zeta)|, |(d/d\zeta)\psi_2(\zeta)| = O(\exp \rho_+\zeta)$ *and* $|\psi_1(\zeta)|, |(d/d\zeta)\psi_1(\zeta)| = O(\exp(\rho_+ + \rho_-)\zeta)$ *as* $\zeta \to +\infty$.

Proof. Noting $\varphi_1 = dq_0/d\zeta$ and $\varphi_2 = dp_0/d\zeta$, we see that

$$\frac{\varphi_2(\zeta)}{\varphi_1(\zeta)} = \frac{dp_0}{dq_0} = \frac{1}{2\lambda\gamma} F_0'(0)(1+\gamma_1(q_0)) \quad \text{as } \zeta \to +\infty.$$

Hereafter, $\gamma_i(q_0)$ $(i = 1, 2, \cdots)$ denote some appropriate functions satisfying $\gamma_i(q_0) = O(|q_0|^{\min(\mu,1)})$. Also, we see that

$$2\mu\left(p_0(\zeta) + \frac{\lambda\gamma}{\mu m}\right) = 2\mu\left[\left(p_0(\zeta) + \frac{2\lambda\gamma}{\mu m}\right) \Big/ q_0(\zeta)\right] q_0(\zeta) - \frac{2\lambda\gamma}{m}$$

$$= \frac{\mu}{\lambda\gamma} F_0'(0)(1+\gamma_2(q_0))q_0 - \frac{2\lambda\gamma}{m}.$$

Thus, we have

$$A(\zeta) = \frac{m+1}{2\lambda\gamma\mu} F_0'(0)(1+\gamma_3(q_0))q_0 - \frac{2\lambda\gamma}{m}.$$

Integration of this gives

$$\int_0^\zeta A(\zeta')\, d\zeta' = \frac{m+1}{2\lambda\gamma\mu} F_0'(0) \int_0^\zeta [1+\gamma_3(q_0)]q_0\, d\zeta' - \frac{2\lambda\gamma}{m}\zeta$$

$$= O(1) - \frac{2\lambda\gamma}{m}\zeta$$

so that we have

$$\exp\left(-\int_0^\zeta A(\zeta')\, d\zeta'\right) = O(\exp \rho_+\zeta).$$

Since $\varphi_1 = dq_0/d\zeta = q_0 p_0$ and $p_0 = O(1)$ for $\zeta \geqq 0$, we have

$$\chi(\zeta) = \int_0^\zeta \frac{1}{p_0(\zeta')} \exp\left(-\int_0^{\zeta'} A(\zeta'')\, d\zeta''\right) d\zeta'' = O(\exp \rho_+\zeta).$$

By using these estimates, elementary calculation proves $\psi_1(\zeta) = O(\exp(\rho_+ + \rho_-)\zeta)$ and $\psi_2(\zeta) = O(\exp \rho_+\zeta)$. Differentiating (5.1) with respect to ζ and applying the estimate that $d\varphi_1/d\zeta$, $d\varphi_2/d\zeta = O(\exp \rho_-\zeta)$, we also have $d\psi_1/d\zeta = O(\exp(\rho_+ + \rho_-)\zeta)$ and $d\psi_2/d\zeta = O(\exp \rho_+\zeta)$. This completes the proof.

Using these solutions, we can write the general solution of $(d/d\zeta - B(\zeta))y = k$ with $y = {}^t(y_1, y_2)$ and $k = {}^t(k_1, k_2)$ as

$$\begin{aligned} y(\zeta) = &\left(c_1 + \int_0^\zeta D(\zeta')[\psi_2(\zeta')k_1(\zeta') - \psi_1(\zeta')k_2(\zeta')]\, d\zeta'\right)\varphi(\zeta) \\ &+ \left(c_2 + \int_0^\zeta D(\zeta')[-\varphi_2(\zeta')k_1(\zeta') + \varphi_1(\zeta')k_2(\zeta')]\, d\zeta'\right)\psi(\zeta), \end{aligned} \tag{5.2}$$

where $D(\zeta) = [\det(\varphi(\zeta), \psi(\zeta))]^{-1}$ and c_1, c_2 are arbitrary constants. Since $D(\zeta)^{-1} = \varphi_1\psi_2 - \varphi_2\psi_1 = \varphi_1(\zeta) \exp(-\int_0^\zeta A(\zeta')\, d\zeta') = O(\exp(\rho_- + \rho_+)\zeta)$, it follows from Proposition 5.1 that

$$D(\zeta)\begin{pmatrix} \psi_2(\zeta) & -\psi_1(\zeta) \\ -\varphi_2(\zeta) & \varphi_1(\zeta) \end{pmatrix} = \begin{pmatrix} O(\exp -\rho_-\zeta) & O(1) \\ O(\exp -\rho_+\zeta) & O(\exp -\rho_+\zeta) \end{pmatrix}. \tag{5.3}$$

The boundary condition on $y_1(0)$ determines $c_1 = 0$. Using (5.3), we consider the behavior of y as $\zeta \to +\infty$ for $k \in X^0_\rho \times X^0$ with $\rho = -\rho_- - \delta \geqq 0$ ($\delta > 0$) as follows:

$$\left|\varphi(\zeta)\int_0^\zeta D[\psi_2 k_1 - \psi_1 k_2]\,d\zeta'\right| \leqq \left(C_1\int_0^\zeta e^{-\rho_-\zeta'}\,e^{-\rho\zeta'}|e^{\rho\zeta'}k_1(\zeta')|\,d\zeta' + C_2\int_0^\zeta |k_2(\zeta')|\,d\zeta'\right)e^{\rho_-\zeta}$$

$$\leqq \left[\frac{C_1}{\delta}(e^{\delta\zeta}-1)\|k_1\|_{X^0_\rho} + C_2\zeta\|k_2\|_{X^0}\right]e^{\rho_-\zeta}$$

$$\leqq C_3(\|k_1\|_{X^0_\rho} + \|k_2\|_{X^0})\,e^{-\rho\zeta}.$$

Similarly, we have

$$\left|\int_0^\zeta D[-\varphi_2 k_1 + \varphi_1 k_2]\,d\zeta'\right| \leqq C_4\int_0^\zeta e^{-\rho_+\zeta'}\,e^{-\rho\zeta'}|\,e^{\rho\zeta'}k_1(\zeta')|\,d\zeta' + C_5\int_0^\zeta e^{-\rho_+\zeta'}|k_2(\zeta')|\,d\zeta'$$

$$\leqq C_6(\|k_1\|_{X^0_\rho} + \|k_2\|_{X^0}).$$

These two inequalities show that the uniform boundedness of a solution requires

$$c_2 + \int_0^\infty D(\zeta')[-\varphi_2(\zeta')k_1(\zeta') + \varphi_1(\zeta')k_2(\zeta')]\,d\zeta' = 0. \tag{5.4}$$

Conversely, if (5.4) holds, then the second term in the right-hand side of (5.2) is evaluated as

$$\left|\psi_1(\zeta)\int_\zeta^{+\infty} D[-\varphi_2 k_1 + \varphi_1 k_2]\,d\zeta'\right|$$

$$\leqq C_7\left(\frac{1}{\rho_+ + \rho}\,e^{-(\rho+\rho_+)\zeta}\|k_1\|_{X^0_\rho} + \frac{1}{\rho_+}\,e^{-\rho_+\zeta}\|k_2\|_{X^0}\right)e^{(\rho_+ + \rho_-)\zeta}$$

$$\leqq C_8(\|k_1\|_{X^0_\rho} + \|k_2\|_{X^0})\,e^{-\rho\zeta},$$

$$\left|\psi_2(\zeta)\int_\zeta^{+\infty} D[-\varphi_2 k_1 + \varphi_1 k_2]\,d\zeta'\right| \leqq C_9(\|k_1\|_{X^0_\rho} + \|k_2\|_{X^0}).$$

Hence, (5.4) is sufficient to assure $y \in X^0_\rho \times X^0$, so that we have a unique solution $y(\zeta) \in X^0_{\rho,0} \times X^0$ given by

$$y = \varphi\int_0^\zeta D[\psi_2 k_1 - \psi_1 k_2]\,d\zeta' + \psi\int_\zeta^{+\infty} D[-\varphi_2 k_1 + \varphi_1 k_2]\,d\zeta'.$$

Differentiating this, we easily see that $y \in X^1_{\rho,0} \times X^1$. Thus we have proved Lemma 3.1.

5.2. Proof of Lemma 3.2. We first show that for small $\kappa \in (0, 1)$ and $\varepsilon \geqq 0$, $\mathcal{F} = Q^{-1}g_1$ maps $B_\kappa(0)$ into $B_\kappa(0)$. To see this, we evaluate $g_1(y_1, \varepsilon R_1(q_1), \varepsilon)$ for $y = {}^t(q_1, p_1) \in B_\kappa(0)$ as follows:

$$\|h_{11}(y_1)\|_{X^0_\rho} = \sup_{\zeta \in R_+} |e^{\rho\zeta}p_1(\zeta)q_1(\zeta)| \leqq \|p_1\|_{X^0}\|q_1\|_{X^0_\rho} \leqq \|y\|^2_{Y_\rho},$$

$$\|h_{12}(y_1)\|_{X^0} \leqq \|F_0(q_0+q_1) - F_0(q_0) - F_0'(q_0)q_1\|_{X^0} + \|\mu p_1^2\|_{X^0}$$

$$\leqq C_1\|q_1^{1+\mu'}\|_{X^0} + \mu\|p_1\|^2_{X^0} \leqq C_2\|y_1\|^{1+\mu'}_{Y_\rho},$$

$$|R(q_1)| \leqq \left| \int_0^\zeta (q_0+q_1)^{1+\mu}\zeta' \right| \leqq \int_0^\zeta e^{-\rho(1+\mu)\zeta'}[e^{\rho\zeta'}(q_0+q_1)]^{1+\mu}\, d\zeta'$$
$$\leqq C_3\|q_0+q_1\|_{X_\rho^0}^{1+\mu} \leqq C_4(1+\|y\|_{Y_\rho})^{1+\mu},$$
$$\|h_2(y_1)\|_{X^0} \leqq \left\| \frac{2\lambda}{m}(p_0+p_1)R(q_1) \right\|_{X^0} + \left\| \frac{2\lambda}{m\mu}(q_0+q_1)^{1+\mu} \right\|_{X^0}$$
$$\leqq C_5(1+\|y\|_{Y_\rho})^{2(1+\mu)} + C_6(1+\|y\|_{Y_\rho})^{1+\mu}.$$

Combining these estimates, we have

$$\|Q_1^{-1}g_1\|_{Y_\rho} \leqq K_1[\|y\|_{Y_\rho}^2 + C_2\|y_1\|_{Y_\rho}^{1+\mu'} + \varepsilon C_7(1+\|y\|_{Y_\rho})^{2(1+\mu)}]$$
$$\leqq K_1[\kappa^2 + C_2\kappa^{1+\mu'} + \varepsilon C_7(1+\kappa)^{2(1+\mu)}],$$

where $K_1 = \|Q_1^{-1}\|_{Y_\rho \to Y_\rho}$.

Thus, we choose κ and $\varepsilon_0(\kappa)$ as $\kappa^{\mu'} \leqq \kappa_1 = 1/2K_1(1+C_2)$ and $\varepsilon_1 = \kappa/2^{3+2\mu}K_1C_7$, so that $\|\mathcal{F}\|_{Y_\rho} \leqq \kappa$ for all $\varepsilon \in [0, \varepsilon_1]$.

Next we consider the contracting property. We easily have

$$\|\mathcal{F}(y_1, \varepsilon) - \mathcal{F}(y_2, \varepsilon)\|_{Y^\rho} \leqq K_1(\|h_{11}(y_1) - h_{11}(y_2)\|_{X_\rho^0} + \|(h_{12}(y_1) - h_{12}(y_2)\|_{X^0})$$
$$+ \varepsilon K_1\|h_{22}(y_1) - h_{22}(y_2)\|_{X^0},$$

where $y_i = {}^t(q_i, p_i)$ for $i = 1, 2$. The right-hand side is estimated as follows:

$$\|h_{11}(y_1) - h_{11}(y_2)\|_{X_\rho^0} \leqq \|(p_1-p_2)q_1\|_{X_\rho^0} + \|p_2(q_1-q_2)\|_{X_\rho^0}$$
$$\leqq \|q_1\|_{X_\rho^0}\|p_1-p_2\|_{X^0} + \|p_2\|_{X^0}\|q_1-q_2\|_{X_\rho^0}$$
$$\leqq \kappa\|y_1-y_2\|_{Y^\rho},$$

$$\|h_{12}(y_1) - h_{12}(y_2)\|_{X^0}$$
$$\leqq \|F_2(q_0+q_1) - F_0(q_0+q_2) - F_0'(q_0)(q_1-q_2)\|_{X^0} + \mu\|p_1^2 - p_2^2\|_{X^0}$$
$$\leqq \|F_0'(q_0+q_1+\theta(q_2-q_1)) - F_0'(q_0)\|_{X^0}\|q_1-q_2\|_{X_\rho^0} + 2\mu\kappa\|p_1-p_2\|_{X^0}$$
$$\leqq (C_8(2\kappa)^{\mu'} + 2\mu\kappa)\|y_1-y_2\|_{Y_\rho},$$

where θ is an appropriate value in $(0, 1)$ and C_8 is the Hölder constant of the first derivative of F_0. Since

$$|R(q_1) - R(q_2)| = \int_0^\zeta |(q_0+q_1)^{1+\mu} - (q_0+q_2)^{1+\mu}|\, d\zeta'$$
$$\leqq (1+\mu)\int_0^\zeta |q_0+q_1+\theta(q_2-q_1)|^\mu\, e^{-\rho\zeta'}\, e^{\rho\zeta'}|q_1-q_2|\, d\zeta'$$
$$\leqq C_9\|q_1-q_2\|_{X_\rho^0},$$

we have

$$\|h_{22}(y_1) - h_{22}(y_2)\|_{X^0}$$
$$\leqq C_{10}(\|R(q_1) - R(q_2)\|_{X^0} + \|p_1R(q_1) - p_2R(q_2)\|_{X^0}$$
$$+ \|(q_0+q_1)^{1+\mu} - (q_0+q_2)^{1+\mu}\|_{X^0})$$
$$\leqq C_{11}(\|q_1-q_2\|_{X_\rho^0} + \|p_1-p_2\|_{X^0}) = C_{11}\|y_1-y_2\|_{Y_\rho^0}.$$

It follows from these estimates that

$$\|\mathcal{F}(y_1, \varepsilon) - \mathcal{F}(y_2, \varepsilon)\|_{Y_\rho^0} \leqq (C_{12}\kappa^{\mu'} + C_{11}K_1\varepsilon)\|y_1-y_2\|_{Y_\rho^0}. \tag{5.5}$$

Hence for $\kappa^{\mu'} = \min(\kappa_1, 1/4C_{12})$ we choose $\varepsilon_0(\kappa) = \min(\varepsilon_1, 1/4C_{11}K_1)$. Then for all $\varepsilon \in [0, \varepsilon_0(\kappa)]$, $\mathcal{F}$ is a contracting mapping. Inequality (5.5) also ensures the uniform continuity of $\mathcal{F}(y_1, \varepsilon)$ with respect to y_1 in $(y_1, \varepsilon) \in B_\kappa(0) \times [0, \varepsilon_0(\kappa)]$. The form of $\mathcal{F}(y_1, \varepsilon)$ directly proves its uniform continuity with respect to ε. This completes the proof.

Acknowledgment. The first author acknowledges the hospitality of the Centre for Mathematical Biology at Oxford University, where part of this work was completed, and its director, Professor J. D. Murray.

REFERENCES

[1] W. ALT, *Contraction patterns in a various polymer system*, in Proc. Modelling of Patterns in Space and Time, W. Jäger and J. D. Murray, eds., Lecture Notes in Biomathematics 55, Springer-Verlag, Berlin, New York, 1984, pp. 1-12.

[2] D. G. ARONSON, *Density dependent interaction-diffusion systems*, in Dynamics and Modelling of Reactive Systems, Academic Press, New York, 1980, pp. 161-176.

[3] D. G. ARONSON, M. G. CRANDALL, AND L. A. PELETIER, *Stabilization of solutions of degenerate nonlinear diffusion problem*, Nonlinear Anal., 6 (1982), pp. 1001-1022.

[4] S. N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[5] E. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[6] S. DUNBAR, *Travelling wave solutions of diffusive Lotka-Volterra equations*, J. Math. Biol., 17 (1983), pp. 11-32.

[7] P. C. FIFE, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497-521.

[8] W. S. C. GURNEY AND R. M. NISBET, *The regulation of inhomogeneous populations*, J. Theoret. Biol., 52 (1975), pp. 441-457.

[9] Y. HOSONO, *Traveling wave solutions for some density dependent diffusion equations*, Japan J. Appl. Math., 3 (1986), pp. 163-196.

[10] ———, *Traveling waves for some biological systems with density dependent diffusion*, Japan J. Appl. Math., 4 (1987), pp. 279-359.

[11] T. IKEDA, *Standing pulse-like solutions of a spatially aggregating population model*, Japan J. Appl. Math., 2 (1985), pp. 111-149.

[12] T. IKEDA AND T. NAGAI, *Stability of localized stationary solutions*, Japan J. Appl. Math., 4 (1987), pp. 73-97.

[13] M. MIMURA, D. TERMAN, AND T. TSUJIKAWA, *Nonlocal advection effect on bistable reaction-diffusion equations*, in Patterns and Waves—Qualitative Analysis of Nonlinear Differential Equations, T. Nishida, M. Mimura, and H. Fujii, eds., North-Holland, Amsterdam, 1986, pp. 507-542.

[14] T. NAGAI AND M. MIMURA, *Some nonlinear degenerate diffusion equations related to population dynamics*, J. Math. Soc. Japan, 35 (1983), pp. 539-562.

[15] ———, *Asymptotic behavior for a nonlinear degenerate diffusion equation in population dynamics*, SIAM J. Appl. Math., 43 (1983), pp. 449-464.

[16] ———, *Asymptotic behavior of the interface to a nonlinear degenerate diffusion equation in population dynamics*, Japan J. Appl. Math., 3 (1986), pp. 129-161.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 870-885, July 1989

006

SMALL PARAMETERS IN STRUCTURED POPULATION MODELS AND THE TROTTER–KATO THEOREM*

H. J. A. M. HEIJMANS† AND J. A. J. METZ‡

Abstract. The justification of some (often implicit) limit arguments used in the development of structured population models is discussed via two examples. The first example shows how a pair of sink-source terms may transform into a side condition relating the appearance of individuals in the interior of the individual state space to the outflow of individuals at its boundary. The second example considers the usual equation for size-dependent population growth in which it is implicitly assumed that discrete finitely-sized young are produced from infinitesimal contributions by all potential parents. The main mathematical tool for dealing with these examples is the Trotter–Kato theorem for one-parameter semigroups of bounded linear operators.

Key words. structured population, limit transition, C_0-semigroup, Trotter–Kato theorem

AMS(MOS) subject classifications. 92A15, 35A35, 47D05

1. Introduction.

1.1. Biological motivation: structured populations, semigroups of operators, and the need for model simplifications. The tenet of the *physiologically structured* approach to the modeling of the dynamics of populations as set out in Metz and Diekmann (1986) is that, provided all individuals experience the same environmental inputs such as food availability or chance of running into a predator, we may (and should) represent a population as a frequency distribution over a space Ω of potential states of the individuals comprising the population. (As we frequently need corresponding concepts on the individual and population levels we will, where necessary, use the prefixes *i*- and *p*- to distinguish the corresponding terms, for example *i*-state versus *p*-state, where the latter refers to the frequency distribution.) The main effort in model construction is the determination of an appropriate state representation of *i*-behavior, where the *i*-behavior consists of (i) any contributions to population change such as giving birth or dying, and (ii) any quantities relevant to the calculation of the output from the population model, such as the rate at which the individual consumes food. If we make the assumption that the number of individuals is sufficiently large, then for any given course of the environment the present *p*-state should determine the future *p*-states in a deterministic and linear fashion. For a constant environment the maps relating subsequent *p*-states should form a linear semigroup.

The transition from *i*-model to *p*-model is made through their differential generators. It is here that we leave biology and start doing mathematics: did we really write down a genuine differential generator, and what can be said about the properties of the semigroup so generated?

Until now the attention has been mostly restricted to models where the *i*-state space Ω is a subset of $\mathbb{R}^k$, and where the individuals move through Ω according to the solution of an ordinary differential equation (ODE), possibly alternating with (usually randomly occurring) state jumps, for example, due to an individual losing weight when it splits off a daughter. The reasons for this restriction are twofold. First, models allowing continuous random *i*-state movements contain many more coefficient functions, which are difficult to specify on a mechanistic basis starting from known

* Received by the editors October 1, 1986; accepted for publication (in revised form) October 4, 1988.
† Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, the Netherlands.
‡ Institute of Theoretical Biology, University of Leiden, Groenhovenstraat 5, 2311 BT Leiden, the Netherlands.

underlying biology. Second, it is generally easier to obtain biological information from less complex models. After all, the goal of the whole exercise is gaining, preferably quantitative, insight into the relation between concrete, though possibly idealized, mechanisms operating in the individuals and consequent population dynamical phenomena. In fact random state jumps are already a bit of a nuisance in that they preclude the sort of simple calculations that a practicing biologist can perform all by himself.

In the present paper we will consider the systematic simplification of two models that both contain random i-state jumps. In the first model, which derives from cell kinetics, we will remove the random character of the jumps by concentrating the takeoffs at one place in Ω only. In the second model, which derives from the population dynamics of ectothermic animals (compare Sinko and Streifer (1967); Streifer (1974); Murphy (1983); Metz, de Roos, and van den Bosch (1988); and in particular Metz and Diekmann (1986)) and also from the dynamics of fungal pellet cultures (compare Edelstein and Hadar (1983); and Chipot and Edelstein (1983)), we will let the size of the jumps become infinitesimally small, while at the same time increasing their occurrence rate. The mathematical tools we use to justify the limit transitions are derived from the theory of one-parameter semigroups of operators (see Pazy (1983)). Particularly important is the Trotter-Kato theorem, which relates the convergence of a sequence of infinitesimal generators (respectively, their resolvents) to the convergence of the associated semigroups. The resulting limit models both allow simple alternative representations in the form of renewal equations for the rates at which newborns appear into the population with kernels, which can easily be calculated in terms of the model ingredients, making possible the routine calculation of biologically relevant quantities such as the asymptotic rate of population increase. Moreover, the limit models contain a smaller number of coefficient functions, making it easier to calibrate them against experimental data.

1.2. Simplification procedures in two special models. In both models considered in this paper the i-state variable of interest will be size; the growth rate of an individual of size x will be denoted as $g(x)$, and Ω will be an interval of $\mathbb{R}_+$. The rate at which individuals of size x die will be denoted as $\mu(x)$.

In the first family of models we consider cells that divide into two at a rate $b_\varepsilon(x)$, where $\varepsilon > 0$ is a small parameter. It is assumed that cells that have passed size one are no longer capable of dividing, but either differentiate or die, i.e., $b_\varepsilon(x) = 0$ for $x > 1$. We will moreover assume that $b_\varepsilon(x) = 0$ for $x < 1 - \varepsilon$. The two daughter cells may differ in size, but the distribution of their relative sizes is constant. This distribution is represented by the probability density $d(p)$, $d : (0, 1) \to \mathbb{R}_+$, where p is the fractional size of the daughter relative to that of its mother. As the sizes of the two daughters add up to the size of the mother, d is symmetric around $\frac{1}{2}$. We will moreover assume that $d(p) = 0$ outside $(\frac{1}{2} - \Delta, \frac{1}{2} + \Delta)$. Finally we assume that ε is so small that the size of the largest newborn daughter is less than the size of the smallest mother, i.e., $\frac{1}{2} + \Delta < 1 - \varepsilon$. Then the size of the smallest daughter, $x_{\min}$, satisfies

$$x_{\min} = (1 - \varepsilon)(\tfrac{1}{2} - \Delta) > (\tfrac{1}{2} + \Delta)(\tfrac{1}{2} - \Delta) =: \alpha.$$

This allows us to choose Ω to be $[\alpha, 1]$ independent of ε. The growth rate g is assumed to be positive and continuous on Ω. Let $n(t, \cdot)$ denote the density function of the cell sizes present at time t; then

$$\text{(1.1a)} \quad \frac{\partial}{\partial t} n(t, x) = -\frac{\partial}{\partial x}(g(x) n(t, x)) - b_\varepsilon(x) n(t, x) + 2 \int_0^1 \frac{d(p)}{p} b_\varepsilon\left(\frac{x}{p}\right) n\left(t, \frac{x}{p}\right) dp,$$

$$\text{(1.1b)} \quad n(t, \alpha) = 0.$$

Now assume that when we let $\varepsilon \downarrow 0$, the quantity

$$\int_{1-\varepsilon}^{1} \frac{b_\varepsilon(y)}{g(y)}\,dy$$

converges to a number larger than zero. This means that the probability that a newborn cell is eventually going to divide

$$\pi_\varepsilon = 1 - \exp\left(-\int_{1-\varepsilon}^{1} \frac{b_\varepsilon(y)}{g(y)}\,dy\right) \tag{1.2}$$

converges to a value $\pi_0 > 0$. In the limit cells will only divide on reaching $x = 1$, and they do so with probability π_0. The corresponding population equation is

$$\frac{\partial}{\partial t} n(t, x) = -\frac{\partial}{\partial x}(g(x)n(t, x)) + 2d(x)\pi_0 g(1)n(t, 1), \tag{1.3a}$$

$$n(t, \alpha) = 0. \tag{1.3b}$$

This limit model may be used as a convenient approximation when cell division occurs only in a narrow size window.

In § 2 we show that under suitable assumptions on the functions g, d, and b_ε, the solutions of (1.1) on the space $L^1[\alpha, 1]$ indeed converge to the solutions of (1.3).

In the second family of models we consider individuals that reproduce at a rate $\beta_\varepsilon(x)$ by splitting off young with size ε, ε small, while concurrently their size is decreased by the same amount. We assume that newborns have viability ρ_ε due to the necessity to survive an infinitesimally short larval stage. The corresponding population equation reads as follows:

$$\frac{\partial n}{\partial t}(t, x) + \frac{\partial}{\partial x}(g(x)n(t, x)) = -\beta_\varepsilon(x)n(t, x) + \beta_\varepsilon(x+\varepsilon)n(t, x+\varepsilon) - \mu(x)n(t, x), \tag{1.4a}$$

$$g(\varepsilon+)n(t, \varepsilon+) - g(\varepsilon-)n(t, \varepsilon-) = \rho_\varepsilon \int_0^1 \beta_\varepsilon(x)n(t, x)\,dx, \tag{1.4b}$$

$$g(0)n(t, 0) = 0, \tag{1.4c}$$

$$n(0, x) = \psi(x). \tag{1.4d}$$

In (1.4b),

$$g(\varepsilon+)n(t, \varepsilon+) - g(\varepsilon-)n(t, \varepsilon-) = \lim_{h \downarrow 0} [g(\varepsilon+h)n(t, \varepsilon+h) - g(\varepsilon-h)n(t, \varepsilon-h)].$$

It is assumed that growth stops at $x = 1$, i.e., $g(1) = 0$, and that g is positive for all smaller sizes including zero. Although the model structure is still compatible with representing the population state as a density function $n(t, \cdot)$, the *jump condition* (1.4b) makes the problem technically troublesome. A natural way out of this dilemma is provided by the observation that the only interesting quantities to be derived from a structured population model are population averages such as total population size, total biomass, or population feeding rate, i.e., linear functionals of $n(t, \cdot)$. This brings us to consider the so-called backward equation

$$\begin{aligned}\frac{\partial m}{\partial t}(t, x) - g(x)\frac{\partial m}{\partial x}(t, x) \\ = -\beta_\varepsilon(x)m(t, x) + \beta_\varepsilon(x)m(t, x-\varepsilon) + \rho_\varepsilon\beta_\varepsilon(x)m(t, \varepsilon) - \mu(x)m(t, x),\end{aligned} \tag{1.5a}$$

$$m(0, x) = \phi(x), \tag{1.5b}$$

satisfied by the clan averages

$$m(t,x)=\int_0^1 \phi(\xi)N_x(t,d\xi), \tag{1.6}$$

where the Borel measure $N_x(t,\cdot)$ represents the expected state at time t of a clan descending from an ancestral individual sized x at time zero. If $m_\varepsilon(t,x;\phi)$ is the solution of (1.5), then every p-output is of the form $\int_0^1 m_\varepsilon(t,x;\phi)\psi(dx)$, where ψ is the initial condition in (1.4d). In § 3.3 we give a precise description of the duality relation between solutions of the forward and backward equations in terms of semigroups and generators (also see Heijmans (1984) and Clement et al. (1987)).

In nature usually roughly the same amount of energy is available for reproduction, which, depending on the species, may be spent on producing a few large or many small young. Therefore we set

$$\beta_\varepsilon(x)=\varepsilon^{-1}b(x). \tag{1.7}$$

Moreover, in species with many small young, infant mortality is generally much higher than when the young are large. If recruitment is to stay bounded when we let ε go to zero we have to put

$$\rho_\varepsilon=\varepsilon r. \tag{1.8}$$

Inserting (1.7) and (1.8) into (1.5a) and letting $\varepsilon\downarrow 0$, we obtain

$$\frac{\partial m}{\partial t}(t,x)-(g(x)-b(x))\frac{\partial m}{\partial x}(t,x)=rb(x)m(t,0)-\mu(x)m(t,x), \tag{1.9}$$

which corresponds to the forward equation

$$\begin{aligned}&\frac{\partial}{\partial t}n(t,x)=-\frac{\partial}{\partial x}(\gamma(x)n(t,x))-\mu(x)n(t,x),\\ &\gamma(x_0)n(t,x_0)=\int rb(y)n(t,y)\,dy,\end{aligned} \tag{1.10}$$

with

$$\gamma(x)=g(x)-b(x) \tag{1.11}$$

and $x_0=0$. Instead of being set back in size at each discrete reproductive event the individual's growth rate is reduced by an amount related to the energy spent in reproduction. Note that in contrast to g the reduced growth rate γ is no longer positive on $[0,1)$, in particular $\gamma(1)=-b(1)<0$.

In § 3.2 we show that under suitable assumptions on the functions g, b, and μ, the semigroup generated by (1.5) indeed converges to the semigroup generated by (1.9). In that section we will also discuss in somewhat more detail the relation between the forward and backward equations.

Equation (1.10) is the equation usually encountered in the population dynamical literature; only x_0 is generally assumed to be positive. Biologically this amounts to the assumption that either parents can time and again produce instantaneously additional masses x_0, notwithstanding the fact that they can add to their own body mass only in a continuous fashion, or else that live newborns are created by magic out of the added infinitesimal contributions by all parents together. Both assumptions go against the grain. Our limiting procedure provides a possible justification, provided x_0 is vanishingly small.

Note for the biological reader. There remains the seemingly awkward assumption that $g(0)>0$. However, the most often encountered biological growth law, the Von Bertalanffy equation, has precisely this property. Note that the Von Bertalanffy Ansatz does not allow individuals to spontaneously spring into being by growing away from size zero. What matters is that $\lim_{x\downarrow 0} g(x)>0$. Individuals of size zero never exist, only individuals that are very small.

Remark. Another way to guarantee that recruitment stays bounded for $\varepsilon\downarrow 0$ is to keep ρ_ε constant and to replace the usual integrability assumption on μ by the assumption that

$$\frac{\mu(x)}{g(x)}=\frac{1}{x}+f(x)$$

with f an L^1 function. To see that this has indeed the intended effect, observe that the probability that a recruited individual survives until it reaches size $x>\varepsilon$ equals

$$\exp\left(-\int_\varepsilon^x \frac{\mu(y)}{g(y)}\,dy\right).$$

(Note that any other choice for the behavior of $\mu(x)$ near $x=0$ does not for $\varepsilon\downarrow 0$ yield the needed survival proportional to ε during the first moments after recruitment!)

2. From distributed to concentrated division.

2.1. The equation and the associated semigroup. In this section we make a thorough mathematical study of (1.1) describing a size-structured cell population reproducing by division. For the sake of convenience we recall the following equation:

$$\text{(2.1a)}\quad \frac{\partial n}{\partial t}(t,x)+\frac{\partial}{\partial x}(g(x)n(t,x))=-b_\varepsilon(x)n(t,x)+2\int_0^1 \frac{d(p)}{p}b_\varepsilon\left(\frac{x}{p}\right)n\left(t,\frac{x}{p}\right)dp,$$

$$\text{(2.1b)}\quad n(t,\alpha)=0,$$

$$\text{(2.1c)}\quad n(0,x)=\phi(x).$$

We will prove that under the right set of assumptions solutions of this problem converge for $\varepsilon\downarrow 0$ to solutions of the limit equation (1.3), i.e.,

$$\text{(2.2a)}\quad \frac{\partial n}{\partial t}(t,x)+\frac{\partial}{\partial x}(g(x)n(t,x))=2\pi_0 d(x)g(1)n(t,1),$$

$$\text{(2.2b)}\quad n(t,a)=0,$$

$$\text{(2.2c)}\quad n(0,x)=\phi(x).$$

We refer to § 1.2 for the interpretation of ε, g, b_ε, d, α, and π_0. As the underlying population state space we choose $X=L^1[\alpha,1]$. We make the following assumptions.

Assumption 2.1. (a) $g\in C[\alpha,1]$; $g(x)>0$, $x\in[\alpha,1]$.

(b) $d\in C[0,1]$; $d(p)>0$ if and only if $|p-\frac{1}{2}|<\Delta$; d is symmetric around $p=\frac{1}{2}$, and $\int_{1/2-\Delta}^{1/2+\Delta} d(p)\,dp=1$.

(c) $b_\varepsilon\in C[\alpha,1]$; $b_\varepsilon(x)=0$, $x\in[\alpha,1-\varepsilon]$; $b_\varepsilon(x)>0$, $x\in(1-\varepsilon,1]$.

We can write (2.1) with initial condition $\phi\in X$ as the abstract Cauchy problem

$$\text{(2.3)}\quad \frac{dn}{dt}(t)=A_\varepsilon n(t),\qquad n(0)=\phi,$$

where the closed operator A_ε on X is given by

$$\text{(2.4)}\quad (A_\varepsilon\phi)(x)=-\frac{d}{dx}(g(x)\phi(x))-b_\varepsilon(x)\phi(x)+2\int_{1/2-\Delta}^{1/2+\Delta}\frac{d(p)}{p}b_\varepsilon\left(\frac{x}{p}\right)\phi\left(\frac{x}{p}\right)dp,$$

for any ϕ in its domain

$$D(A_\varepsilon)=\{\phi\in X:\ g\phi\in W^{1,1}[\alpha,1] \text{ and } \phi(\alpha)=0\}. \tag{2.5}$$

Using a standard perturbation result for C_0-semigroups (Pazy (1983, § 3.1)) we easily show that A_ε is the infinitesimal generator of a strongly continuous semigroup $\{T_\varepsilon(t)\}_{t\geqq 0}$.

Let A be a closed linear operator on the Banach space X and let $M\geqq 0$, $\omega\in\mathbb{R}$. We say that $A\in G(M,\omega)$ if A is the infinitesimal generator of a C_0-semigroup $\{T(t)\}_{t\geqq 0}$ of bounded linear operators satisfying

$$\|T(t)\|\leqq M e^{\omega t},\qquad t\geqq 0$$

(e.g., Pazy (1983, § 3.4)). The next proposition states, among other things, that there exists a semigroup solution to (2.1).

THEOREM 2.2. *There exist constants $\omega\in\mathbb{R}$ and $M\geqq 1$ (which do not depend on ε), such that $A_\varepsilon\in G(M,\omega)$.*

Proof. Let $\|\cdot\|$ be the L^1-norm. Then the norm $\|\cdot\|$ is equivalent to the norm $\|\cdot\|'$ given by

$$\|\phi\|'=\int_\alpha^1 x|\phi(x)|\,dx,\qquad \phi\in X.$$

Let, for $t\geqq 0$,

$$\|T_\varepsilon(t)\|'=\sup\{\|T_\varepsilon(t)\phi\|'/\|\phi\|':\ \phi\in X,\ \phi\neq 0\}.$$

Since $T_\varepsilon(t)$ is a positive operator, we have

$$\|T_\varepsilon(t)\|'=\{\|T_\varepsilon(t)\phi\|'/\|\phi\|':\ \phi\in X_+,\ \phi\neq 0\},$$

where X_+ is the cone of positive elements. If $\phi\in X_+$, then $\|T_\varepsilon(t)\phi\|'=\int_\alpha^1 xn(t,x)\,dx$, where $n(t,x)$ is the solution of (2.1). If, in addition, $\phi\in D(A_\varepsilon)$, then

$$\frac{d}{dt}\int_\alpha^1 xn(t,x)\,dx\leqq\int_\alpha^1 g(x)n(t,x)\,dx\leqq\omega\int_\alpha^1 xn(t,x)\,dx,$$

where $\omega>0$ is taken so large that $g(x)\leqq\omega x$, $x\in[\alpha,1]$. So for $\phi\in D(A_\varepsilon)\cap X_+$ we find that

$$\|T_\varepsilon(t)\phi\|'=\int_\alpha^1 xn(t,x)\,dx\leqq e^{\omega t}\int_\alpha^1 x\phi(x)\,dx=e^{\omega t}\|\phi\|'.$$

Since $D(A_\varepsilon)\cap X_+$ is norm-dense in X_+, this holds for any $\phi\in X_+$, and we find that

$$\|T_\varepsilon(t)\|'\leqq e^{\omega t},\qquad t\geqq 0.$$

Since $\|\cdot\|'$ and $\|\cdot\|$ are equivalent norms, there exists a constant $M>0$ such that

$$\|T_\varepsilon(t)\|\leqq M e^{\omega t},\qquad t\geqq 0,$$

and the result is proved. □

2.2. Justification of the limit transition. In this section we give a formal mathematical justification of the limit transition $\varepsilon\downarrow 0$ which amounts to (2.2). That is to say, we prove that the solution of (2.1) given by $n(t,\cdot)=T_\varepsilon(t)\phi$ converges to the solution of (2.2) as $\varepsilon\downarrow 0$. For this purpose, we use the Trotter–Kato theorem. Besides Assumptions 2.1(a)–(c) we only assume there exists a $\pi_0\in[0,1)$ such that $\lim_{\varepsilon\downarrow 0}\pi_\varepsilon=\pi_0$.

We rewrite (2.2) as the abstract Cauchy problem

$$\frac{dn}{dt}(t) = An(t), \qquad n(0) = \phi, \tag{2.6}$$

where A is the closed operator

$$(A\phi)(x) = -\frac{d}{dx}(g(x)\phi(x)) + 2\pi_0 d(x) g(1)\phi(1) \tag{2.7}$$

with dense domain

$$D(A) = \{\phi \in X: g\phi \in W^{1,1}[\alpha, 1] \text{ and } \phi(\alpha) = 0\}. \tag{2.8}$$

THEOREM 2.3. *For $\lambda \in \mathbb{R}$ large enough we have*

$$R(\lambda, A_\varepsilon)\phi \to R(\lambda, A)\phi, \qquad \varepsilon \downarrow 0,$$

for every $\phi \in X$.

Proof. The proof consists of four steps.

(1) Let the isomorphism $U_\varepsilon: X \to X$ be given by

$$(U_\varepsilon \phi)(x) = \frac{E_\varepsilon(x)}{g(x)}\phi(x),$$

where $E_\varepsilon(x) = \exp\left(-\int_\alpha^x (b_\varepsilon(y)/g(y))\, dy\right)$. Let

$$D = D(A) = D(A_\varepsilon) = \{\phi \in X: g\phi \in W^{1,1}[\alpha, 1] \text{ and } \phi(\alpha) = 0\},$$

and

$$\tilde{D} = U_\varepsilon^{-1} D = \{\phi \in X: \phi \in W^{1,1}[\alpha, 1] \text{ and } \phi(\alpha) = 0\}.$$

Let $\tilde{A}_\varepsilon$ be the closed operator $U_\varepsilon^{-1} A_\varepsilon U_\varepsilon$ with domain $D(\tilde{A}_\varepsilon) = \tilde{D}$. Then $\tilde{A}_\varepsilon$ is given by

$$(\tilde{A}_\varepsilon \phi)(x) = -g(x)\frac{d\phi}{dx}(x) + 2\frac{g(x)}{E_\varepsilon(x)} \int_{1/2-\Delta}^{1/2+\Delta} \frac{d(p)}{p} r_\varepsilon\left(\frac{x}{p}\right)\phi\left(\frac{x}{p}\right) dp,$$

where $r_\varepsilon(x) = (b_\varepsilon(x)/g(x))E_\varepsilon(x)$ for $x \in [\alpha, 1]$. We define the isomorphism $U: X \to X$ by

$$(U\phi)(x) = \frac{\phi(x)}{g(x)}.$$

Let $\tilde{A}$ be the closed operator $U^{-1}AU$ with domain $D(\tilde{A}) = U^{-1}D = \tilde{D}$. For $\phi \in D(\tilde{A})$ we have

$$(\tilde{A}\phi)(x) = -g(x)\frac{d\phi}{dx}(x) + 2\pi_0\, d(x) g(x)\phi(1).$$

(2) We show that for every $\phi \in \tilde{D}$,

$$\tilde{A}_\varepsilon \phi \to \tilde{A}\phi \quad \text{as } \varepsilon \downarrow 0.$$

Let $\phi \in \tilde{D}$, then

$$(\tilde{A}_\varepsilon \phi)(x) - (\tilde{A}\phi)(x) = 2\frac{g(x)}{E_\varepsilon(x)} \int_{1/2-\Delta}^{1/2+\Delta} \frac{d(p)}{p} r_\varepsilon\left(\frac{x}{p}\right)\phi\left(\frac{x}{r}\right) dp - 2\pi_0\, d(x) g(x)\phi(1).$$

We define $\bar{g} = \max_{x\in[\alpha,1]} g(x)$, $\bar{d} = \max_{p\in[1/2-\Delta,1/2+\Delta]} d(p)$. Now

$$\|\tilde{A}_\varepsilon\phi - \tilde{A}\phi\| = \int_\alpha^1 \left| \frac{2g(x)}{E_\varepsilon(x)} \int_{1/2-\Delta}^{1/2+\Delta} \frac{d(p)}{p} r_\varepsilon\left(\frac{x}{p}\right)\phi\left(\frac{x}{p}\right) dp - 2\pi_0\, d(x)g(x)\phi(1) \right| dx$$

$$= \int_{(1/2-\Delta)(1-\varepsilon)}^{1/2+\Delta} \left| 2g(x) \int_{1/2-\Delta}^{1/2+\Delta} \frac{d(p)}{p} r_\varepsilon\left(\frac{x}{p}\phi\left(\frac{x}{p}\right)\right) dp - 2\pi_0\, d(x)g(x)\phi(1) \right| dx$$

$$\leqq 2\bar{g} \int_{(1/2-\Delta)(1-\varepsilon)}^{1/2+\Delta} \left| \int_{1/2-\Delta}^{1/2+\Delta} \frac{d(p)}{p} r_\varepsilon\left(\frac{x}{p}\right)\phi\left(\frac{x}{p}\right) dp - \pi_\varepsilon\, d(x)\phi(1) \right| dx$$

$$+ 2\bar{g}\,|\pi_\varepsilon - \pi_0| \cdot |\phi(1)|.$$

This second expression at the right-hand side can easily be estimated. We write the first expression as the sum of three integrals:

$$\int_{(1/2-\Delta)(1-\varepsilon)}^{1/2+\Delta} = \int_{(1/2-\Delta)(1-\varepsilon)}^{1/2-\Delta} + \int_{1/2-\Delta}^{(1/2+\Delta)(1-\varepsilon)} + \int_{(1/2+\Delta)(1-\varepsilon)}^{1/2+\Delta}.$$

It is the middle integral that causes the most trouble, and we restrict our attention to this term. Let $\delta > 0$:

$$2\bar{g} \int_{1/2-\Delta}^{(1/2+\Delta)(1-\varepsilon)} \left| \int_{1/2-\Delta}^{1/2+\Delta} \frac{d(p)}{p} r_\varepsilon\left(\frac{x}{p}\right)\phi\left(\frac{x}{p}\right) dp - \pi_\varepsilon\, d(x)\phi(1) \right| dx$$

$$= 2\bar{g} \int_{1/2-\Delta}^{(1/2+\Delta)(1-\varepsilon)} \left| \int_{x/(1/2+\Delta)}^{x/(1/2-\Delta)} \frac{1}{y} d\left(\frac{x}{y}\right) r_\varepsilon(y)\phi(y)\, dy - \pi_\varepsilon\, d(x)\phi(1) \right| dx$$

$$= 2\bar{g} \int_{1/2-\Delta}^{(1/2+\Delta)(1-\varepsilon)} \left| \int_{1-\varepsilon}^{1} \frac{1}{y} d\left(\frac{x}{y}\right) r_\varepsilon(y)\phi(y)\, dy - \int_{1-\varepsilon}^{1} d(x) r_\varepsilon(y)\phi(1)\, dy \right| dx$$

$$= 2\bar{g} \int_{1/2-\Delta}^{(1/2+\Delta)(1-\varepsilon)} \left| \int_{1-\varepsilon}^{1} \left\{ \frac{1}{y} d\left(\frac{x}{y}\right)\phi(y) - \frac{1}{1} d\left(\frac{x}{1}\right)\phi(1) \right\} r_\varepsilon(y)\, dy \right| dx$$

$$\leqq 2\bar{g} \int_{1/2-\Delta}^{(1/2+\Delta)(1-\varepsilon)} \left\{ \int_{1-\varepsilon}^{1} \delta r_\varepsilon(y)\, dy \right\} dx \leqq 2\bar{g} \cdot 2\Delta \cdot \delta.$$

Here we have chosen $\varepsilon > 0$ so small that

$$\left| \frac{1}{y} d\left(\frac{x}{y}\right)\phi(y) - \frac{1}{1} d\left(\frac{x}{1}\right)\phi(1) \right| < \delta,$$

for every $x \in [\frac{1}{2}-\Delta, (\frac{1}{2}+\Delta)(1-\varepsilon)]$ and $y \in [1-\varepsilon, 1]$, and we used that $\int_{1-\varepsilon}^{1} r_\varepsilon(y)\, dy = \pi_\varepsilon \leqq 1$. This shows that $\tilde{A}_\varepsilon\phi \to \tilde{A}\phi$ as $\varepsilon \downarrow 0$, for $\phi \in \tilde{D}$.

(3) We show that for $\lambda \in \mathbb{R}$ large enough (in particular $\lambda > \omega$; see Theorem 2.2)

$$R(\lambda, \tilde{A}_\varepsilon)\phi \to R(\lambda, \tilde{A})\phi \quad \text{as } \varepsilon \downarrow 0,$$

for every $\phi \in X$. Choose $\lambda > \omega$ so large that $\lambda \in \rho(A) = \rho(\tilde{A})$. Let $\phi \in X$, and define $\psi \in \tilde{D}$ as $\psi = R(\lambda, \tilde{A})\phi$. For $\varepsilon > 0$, let $\phi_\varepsilon = (\lambda - \tilde{A}_\varepsilon)\psi$. From

$$\tilde{A}_\varepsilon\psi \to \tilde{A}\psi \quad \text{as } \varepsilon \downarrow 0$$

we get $\phi_\varepsilon \to \phi$ as $\varepsilon \downarrow 0$. Since $R(\lambda, \tilde{A}_\varepsilon) = U_\varepsilon^{-1} R(\lambda, A_\varepsilon) U_\varepsilon$, we deduce from Theorem 2.2 that

$$\|R(\lambda, \tilde{A}_\varepsilon)\| \leqq \frac{\tilde{M}}{\lambda - \omega}, \qquad \varepsilon > 0.$$

Here we have used explicitly that $\pi_0 < 1$. Now

$$\lim_{\varepsilon \downarrow 0} R(\lambda, \tilde{A}_\varepsilon)\phi = \lim_{\varepsilon \downarrow 0} [R(\lambda, \tilde{A}_\varepsilon)(\phi - \phi_\varepsilon) + \psi] = \psi = R(\lambda, \tilde{A})\phi.$$

(4) We finally show that for $\lambda \in \mathbb{R}$ large enough,

$$R(\lambda, A_\varepsilon)\phi \to R(\lambda, A)\phi \quad \text{as } \varepsilon \downarrow 0,$$

for every $\phi \in X$. It is easily checked that

$$U_\varepsilon \phi \to U\phi, \quad \varepsilon \downarrow 0 \quad \text{and} \quad U_\varepsilon^{-1}\phi \to U^{-1}\phi, \quad \varepsilon \downarrow 0,$$

for every $\phi \in X$, and that there exists a constant $L > 0$ such that $\|U_\varepsilon\|$, $\|U\|$, $\|U_\varepsilon^{-1}\|$, $\|U^{-1}\| \leqq L$, $\varepsilon > 0$. For every $\phi \in X$ we have

$$\begin{aligned}
\|R(\lambda, A_\varepsilon)\phi - R(\lambda, A)\phi\| &= \|U_\varepsilon R(\lambda, \tilde{A}_\varepsilon) U_\varepsilon^{-1}\phi - UR(\lambda, A)U^{-1}\phi\| \\
&= \|(U_\varepsilon - U)(R(\lambda, \tilde{A}_\varepsilon)U_\varepsilon^{-1} - R(\lambda, \tilde{A}_\varepsilon)U^{-1} + R(\lambda, \tilde{A}_\varepsilon)U^{-1} \\
&\quad - R(\lambda, \tilde{A})U^{-1} + R(\lambda, \tilde{A})U^{-1})\phi \\
&\quad + U(R(\lambda, \tilde{A}_\varepsilon) - R(\lambda, \tilde{A}))(U_\varepsilon^{-1} - U^{-1} + U^{-1})\phi \\
&\quad + UR(\lambda, \tilde{A})(U_\varepsilon^{-1} - U^{-1})\phi\| \\
&\leqq \|U_\varepsilon - U\| \, \|R(\lambda, \tilde{A}_\varepsilon)\| \, \|U_\varepsilon^{-1}\phi - U^{-1}\phi\| \\
&\quad + \|U_\varepsilon - U\| \, \|R(\lambda, \tilde{A}_\varepsilon)U^{-1}\phi - R(\lambda, \tilde{A})U^{-1}\phi\| \\
&\quad + \|(U_\varepsilon - U)R(\lambda, \tilde{A})U^{-1}\phi\| \\
&\quad + \|U\| \, \|R(\lambda, \tilde{A}_\varepsilon) - R(\lambda, \tilde{A})\| \, \|U_\varepsilon^{-1}\phi - U^{-1}\phi\| \\
&\quad + \|U\| \, \|(R(\lambda, \tilde{A}_\varepsilon) - R(\lambda, \tilde{A}))U^{-1}\phi\| \\
&\quad + \|U\| \, \|R(\lambda, \tilde{A})\| \, \|U_\varepsilon^{-1}\phi - U^{-1}\phi\|,
\end{aligned}$$

and all these terms go to zero as $\varepsilon \downarrow 0$. □

We can now apply the Trotter-Kato theorem (Pazy (1983, § 3.4)), which yields that (i) A is the infinitesimal generator of a C_0-semigroup (in particular this means that (2.2) is well posed) and that (ii) the solution of (2.1) converges to the solution of (2.2) as $\varepsilon \downarrow 0$.

THEOREM 2.4. *$A \in G(M, \omega)$, and if $\{T(t)\}_{t \geqq 0}$ is the semigroup generated by A, then for every $\phi \in X$, $t \geqq 0$,*

$$T_\varepsilon(t)\phi \to T(t)\phi \quad \text{as } \varepsilon \downarrow 0.$$

Moreover, the convergence is uniform with respect to t in bounded subsets of $(0, \infty)$.

3. From size jumps to reduced growth.

3.1. The semigroup solution to the backward equation. In this section we show that under some reasonable assumptions we can associate a C_0-semigroup of bounded linear operators on $X = C[0, 1]$ with the backward equation (1.5), which we recall below for convenience. Throughout this section we will assume that the death rate μ

is identically zero. However, all the results obtained here remain valid for nonzero death rates. The backward equation reads as follows:

$$\frac{\partial m}{\partial t}(t,x)-g(x)\frac{\partial m}{\partial x}(t,x)=-\beta_\varepsilon(x)n(t,x)+\beta_\varepsilon(x)m(t,x-\varepsilon)+\rho_\varepsilon\beta_\varepsilon(x)m(t,\varepsilon), \tag{3.1a}$$

$$m(0,x)=\phi(x). \tag{3.1b}$$

Assumption 3.1. (a) g is Lipschitz continuous on $[0,1]$; $g(x)>0$, $x\in[0,1)$; $g(1)=0$.

(b) β_ε is Lipschitz continuous on $[0,1]$; there is an $a>\varepsilon$ such that $\beta_\varepsilon(x)=0$, $x\in[0,a]$ and $\beta_\varepsilon(x)>0$, $x\in(a,1]$.

Here a denotes the minimum size at which an individual can reproduce. We can write (3.1) as the abstract Cauchy problem:

$$\frac{dm}{dt}(t)=A_\varepsilon m(t), \tag{3.2a}$$

$$m(0)=\phi\in X, \tag{3.2b}$$

where the closed unbounded operator A_ε with domain

$$D(A_\varepsilon)=\{\phi\in X\cap W^{1,1}_{\mathrm{loc}}[0,1]\colon g\phi'\in X\},$$

is given by

$$(A_\varepsilon\phi)(x)=g(x)\frac{d\phi}{dx}(x)-\beta_\varepsilon(x)\phi(x)+\beta_\varepsilon(x)\phi(x-\varepsilon)+\rho_\varepsilon\beta_\varepsilon(x)\phi(\varepsilon).$$

We write A_ε as the sum of two operators:

$$A_\varepsilon=A_0+B_\varepsilon, \tag{3.3}$$

where the closed unbounded operator A_0 has the same domain as A_ε and is given by

$$(A_0\phi)(x)=g(x)\frac{d\phi}{dx}(x),$$

and where B_ε is a bounded operator given by

$$(B_\varepsilon\phi)(x)=-\beta_\varepsilon(x)\phi(x)+\beta_\varepsilon(x)\phi(x-\varepsilon)+\rho_\varepsilon\beta_\varepsilon(x)\phi(\varepsilon).$$

It is quite easy to show that A_0 generates a strongly continuous semigroup $\{T_0(t)\}_{t\geqq 0}$, and therefore A_ε, being a bounded perturbation of A_0, also generates a strongly continuous semigroup $\{T_\varepsilon(t)\}_{t\geqq 0}$ (see Pazy (1983, § 3.1)).

Both $\{T_0(t)\}_{t\geqq 0}$ and $\{T_\varepsilon(t)\}_{t\geqq 0}$ are positive semigroups, which is intuitively clear from the biological interpretation, but can also be shown rigorously (see Heijmans (1986)). Let $\mathbf{1}$ be the element of X that is identically one on $[0,1]$. Then

$$A_\varepsilon\mathbf{1}=\rho_\varepsilon\beta_\varepsilon.$$

Define the positive scalar ω_ε by

$$\omega_\varepsilon=\sup\{\rho_\varepsilon\beta_\varepsilon(x)\colon x\in[0,1]\}. \tag{3.4}$$

We see immediately that

$$0\leqq A_\varepsilon\mathbf{1}\leqq\omega_\varepsilon\mathbf{1}.$$

We show that $A_\varepsilon \in G(1, \omega_\varepsilon)$. First suppose that $\omega_\varepsilon < s(A_\varepsilon)$, where $s(A_\varepsilon)$ is the spectral bound of A_ε, i.e., $s(A_\varepsilon) = \sup\{\mathrm{Re}\,\lambda : \lambda \in \sigma(A_\varepsilon)\}$. Since $\{T_\varepsilon(t)\}_{t\geqq 0}$ is a positive semigroup, $s(A_\varepsilon) \in \sigma(A_\varepsilon)$ if $\sigma(A_\varepsilon) \neq \emptyset$, and $R(\lambda, A_\varepsilon)$ is a positive operator if $\lambda > s(A_\varepsilon)$ (see Nagel (1986)). Choose $\lambda > s(A_\varepsilon)$. Since $R(\lambda, A_\varepsilon)$ is a positive operator we get that

$$0 \leqq R(\lambda, A_\varepsilon)\mathbf{1} \leqq \frac{1}{\lambda - \omega_\varepsilon}\mathbf{1};$$

hence $\|R(\lambda, A_\varepsilon)\| = \|R(\lambda, A_\varepsilon)\mathbf{1}\| \leqq 1/(\lambda - \omega_\varepsilon)$, and we find that $\|R(\lambda, A_\varepsilon)\|$ remains bounded if $\lambda \downarrow s(A_\varepsilon)$, which is in contradiction with

$$s(A_\varepsilon) \in \sigma(A_\varepsilon).$$

Therefore $\omega_\varepsilon \geqq s(A_\varepsilon)$. Using the same arguments as above, we find that for $\lambda > \omega_\varepsilon$,

$$\|R(\lambda, A_\varepsilon)\| \leqq \frac{1}{\lambda - \omega_\varepsilon},$$

which yields that for $n = 1, 2, \cdots$

$$\|R(\lambda, A_\varepsilon)^n\| \leqq \frac{1}{(\lambda - \omega_\varepsilon)^n},$$

and it follows from the Hille-Yosida theorem that $A_\varepsilon \in G(1, \omega_\varepsilon)$. In particular this implies that A_ε is the generator of a C_0-semigroup $\{T_\varepsilon(t)\}_{t\geqq 0}$ on X.

3.2. The limit transition justified. Assuming (1.7) and (1.8), i.e., $\beta_\varepsilon(x) = \varepsilon^{-1}b(x)$ and $\rho_\varepsilon = \varepsilon r$, we find the limiting equation

$$\frac{\partial m}{\partial t}(t, x) - \gamma(x)\frac{\partial m}{\partial x}(t, x) = rb(x)m(t, 0), \tag{3.5a}$$

$$m(0, x) = \phi(x), \tag{3.5b}$$

where γ is the *reduced growth rate*

$$\gamma(x) = g(x) - b(x). \tag{3.6}$$

Note that it follows from Assumption 3.1 that (i) b is Lipschitz continuous on $[0, 1]$, $b(x) = 0$ for $x \leqq a$ and $b(x) > 0$ for $a < x \leqq 1$, and that (ii) γ is *not* positive on the whole interval $[0, 1]$, in particular $\gamma(1) = -b(1) < 0$.

In the rest of this section we will show how the Trotter-Kato theorem can be used to justify the formal transition from (3.1)-(3.5). In the next section we interpret these results in terms of the forward equations (1.4) and (1.10) (with $x_0 = 0$).

First we reformulate (3.5a) supplied with initial condition (3.5b) as an abstract Cauchy problem:

$$\frac{dm}{dt}(t) = Am(t), \qquad m(0) = \phi \in X, \tag{3.7}$$

where the closed operator A is given by

$$(A\phi)(x) = \gamma(x)\frac{d\phi}{dx}(x) + rb(x)\phi(0)$$

for every ϕ in its domain

$$D(A) = \{\phi \in X \cap W^{1,1}_{\mathrm{loc}}[0, 1] : \gamma\phi' \in X\}.$$

It is not difficult to show that A generates a strongly continuous positive semigroup: this, however, will also follow from the forthcoming analysis. Let

$$\omega := \{rb(x): 0 \leqq x \leqq 1\}. \tag{3.8}$$

Obviously, $\omega_\varepsilon = \omega$ and from the results of § 3.1 it follows that

$$A_\varepsilon \in G(1, \omega). \tag{3.9}$$

Assumption 3.2. There exists a unique $\hat{x} \in (0, 1)$ such that $g(\hat{x}) = b(\hat{x})$.

In combination with the other assumptions of this section this means that

$$\gamma(x) > 0, \qquad 0 \leqq x < \hat{x},$$
$$\gamma(x) < 0, \qquad \hat{x} < x \leqq 1.$$

Now let $D = C^1[0, 1]$, i.e., the subspace of X containing all continuously differentiable functions on $[0, 1]$. Clearly

$$D(A_\varepsilon) \subseteq D, \qquad D(A) \subseteq D.$$

PROPOSITION 3.3. *$(\lambda - A)D$ is dense in X for $\lambda \in \mathbb{R}$ sufficiently large.*

Proof. Consider for $F \in X$ the inhomogeneous equation

$$\lambda\phi(x) - \gamma(x)\phi'(x) = F(x),$$

where $\lambda \in \mathbb{R}$ is sufficiently large ($\lambda > \omega$). The solution of this equation for $0 \leqq x < \hat{x}$ is given by

$$\phi(x) = \int_x^{\hat{x}} \frac{F(y)}{\gamma(y)} \exp\left\{-\lambda \int_x^y \frac{d\xi}{\gamma(\xi)}\right\} dy, \tag{*}$$

and a similar expression can be found for $\phi(x)$, if x is greater than $\hat{x}$. It is easy to check that $\phi \in D$ if $F \in D$. Now, for $f \in X$, the solution of

$$\lambda\phi - A\phi = f, \tag{**}$$

on $(0, \hat{x})$ is given by (*), with $F(x) = f(x) + r\phi(0)b(x)$ substituted. Hence $\phi \in D$ if $F \in D$. Let $f \in X$ and let ϕ be the solution of (**); then

$$\phi(0) = \int_0^{\hat{x}} \frac{f(y) + r\phi(0)b(y)}{\gamma(y)} \exp\left\{-\lambda \int_0^y \frac{d\xi}{\gamma(\xi)}\right\} dy.$$

We assume that $\lambda \in \mathbb{R}$ is so large that

$$\alpha_\lambda := \int_0^{\hat{x}} \frac{b(y)}{\gamma(y)} \exp\left\{-\lambda \int_0^y \frac{d\xi}{\gamma(\xi)}\right\} dy < \frac{1}{r},$$

and for $f \in X$ we define

$$H_\lambda(f) := \frac{r}{1 - \alpha_\lambda r} \int_0^{\hat{x}} \frac{f(y)}{\gamma(y)} \exp\left\{-\lambda \int_0^y \frac{d\xi}{\gamma(\xi)}\right\} dy.$$

Then the solution ϕ of (**) satisfies

$$r\phi(0) = H_\lambda(f).$$

So we get that $\phi \in D$ if $f + H_\lambda(f)b \in D$. We define $V \subseteq X$ as

$$V = \{f \in X: f + H_\lambda(f)b \in D\}.$$

Then $V \subseteq (\lambda - A)D$, and it suffices to show that V is a dense subset of X. Let $f \in X$ and define $g \in X$ as

$$g = f + H_\lambda(f)b.$$

Let $\{g_n\}$ be a sequence in D converging to g as $n \to \infty$. The solution of

$$g_n = f_n + H_\lambda(f_n)b$$

is given by

$$f_n = g_n - \frac{H_\lambda(g_n)}{1+H_\lambda(b)}\, b.$$

Now $f_n \in V$ and

$$f_n \to g - \frac{H_\lambda(g)}{1+H_\lambda(b)} \cdot b = f,\ n \to \infty.$$

Therefore $\bar{V} = X$. □

PROPOSITION 3.4. *$A_\varepsilon\phi \to A\phi$ as $\varepsilon \downarrow 0$, for every $\phi \in D$.*

Proof. Let $\phi \in D$. Then

$$|(A_\varepsilon\phi)(x) - (A\phi)(x)| \leqq |b(x)| \cdot \left|\frac{1}{\varepsilon}(\phi(x) - \phi(x-\varepsilon)) - \phi'(x)\right| + r|b(x)| \cdot |\phi(\varepsilon) - \phi(0)|,$$

for every $x \in [0, 1]$, and thus

$$\|A_\varepsilon\phi - A\phi\| = \sup_{x\in[0,1]} |(A_\varepsilon\phi)(x) - (A\phi)(x)| \to 0, \qquad \varepsilon \downarrow 0. \qquad \square$$

We are now ready to apply the Trotter-Kato theorem which gives us the following theorem.

THEOREM 3.5. *$A \in G(1, \omega)$, and if $\{T(t)\}_{t\geqq 0}$ is the semigroup generated by A, then*

$$T_\varepsilon(t)\phi \to T(t)\phi, \qquad \varepsilon \downarrow 0,$$

for every $\phi \in X$, where the convergence is uniform for t in bounded subsets of $(0, \infty)$.

This theorem tells us that a solution of the limit equation (3.8) is indeed an approximation of solutions of equation (3.2a), presupposed that their initial condition ϕ is the same.

We can give a very precise description of the relation between the backward and the forward equations and their respective solutions in semigroup terms. It is the backward equation that can be derived rigorously and that is to be solved on the space of continuous functions. Let A_ε be the differential operator on X associated with the backward problem (see (3.2)). Then, by definition, the abstract forward equation is

$$\frac{dn}{dt}(t) = A_\varepsilon^* n(t), \qquad n(0) = \psi \in X^*, \tag{3.10}$$

where A_ε^*, the dual operator of A_ε, is defined on the dual space $X^* = M[0, 1]$, the space of regular Borel measures on $[0, 1]$. The solutions of (3.10) are given by $n_\varepsilon(t, \cdot\,; \psi) = T_\varepsilon^*(t)\psi$. Here the notion of solution must be understood in terms of the weak* topology on X^*. The dual semigroup $\{T_\varepsilon^*(t)\}_{t\geqq 0}$ is a weakly* continuous semigroup with weak* generator A_ε^* (see Butzer and Berens (1967)). There exists the following duality relation between solutions of the forward and the backward equations. For $\phi \in X$ we have

$$\begin{aligned} \int_0^1 \phi(x) n_\varepsilon(t, dx; \psi) &= \langle \phi, n_\varepsilon(t, \cdot\,; \psi)\rangle = \langle \phi, T_\varepsilon^*(t)\psi\rangle \\ &= \langle T_\varepsilon(t)\phi, \psi\rangle = \langle m_\varepsilon(t, \cdot\,; \phi), \psi\rangle = \int_0^1 m_\varepsilon(t, x; \phi)\psi(dx), \end{aligned} \tag{3.11}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between X and X^*, and where $m_\varepsilon(t, \cdot\,; \phi)$ is the solution of the backward problem (3.1).

Let $X^{\odot}$ be the closed subspace of X^* where $\{T_\varepsilon^*(t)\}_{t\geq 0}$ is strongly continuous. Then $X^{\odot} = \overline{D(A_\varepsilon^*)}$ (see Butzer and Berens (1967)). It can be shown (compare the remark below) that in the present situation $X^{\odot} = L^1[0, 1]$ (e.g., Clément et al. (1987); (Clément, Heijmans et al. (1987)). Obviously, $X^{\odot}$ is invariant under $\{T_\varepsilon^*(t)\}_{t\geq 0}$, and the restriction $\{T_\varepsilon^{\odot}(t)\}_{t\geq 0}$ is a strongly continuous semigroup. If we denote its generator by $A_\varepsilon^{\odot}$, then

$$\frac{dn}{dt}(t) = A_\varepsilon^{\odot} n(t), \qquad n(0) = \psi \in X^{\odot},$$

is the abstract formulation of (1.4) and with this observation the circle is closed.

Remark. To prove the latter statement, we have to calculate A_ε^* and its domain $D(A_\varepsilon^*)$ from A_ε and $D(A_\varepsilon)$. This calculation involves the following steps (e.g., Heijmans (1984)):

(i) compute the resolvent operator $R(\lambda, A_\varepsilon)$

(ii) compute its dual $R(\lambda, A_\varepsilon^*) = R(\lambda, A_\varepsilon)^*$

(iii) find the domain $D(A_\varepsilon^*)$ from the relation

$$D(A_\varepsilon^*) = \text{Ran}\,(R(\lambda, A_\varepsilon^*)),$$

Ran$(\cdot)$ denoting the range

(iv) calculate $A_\varepsilon^* \psi$, where $\psi \in D(A_\varepsilon^*)$, from the relation

$$\langle \phi, A_\varepsilon^* \psi \rangle = \langle A_\varepsilon \phi, \psi \rangle \quad \text{for } \phi \in D(A_\varepsilon)$$

(v) $X^{\odot} = \overline{D(A_\varepsilon^*)}$ and $A_\varepsilon^{\odot}$ is the part of A_ε^* in $X^{\odot}$ (e.g., Butzer and Berens (1967)).

Our main result of this section, Theorem 3.5, can be restated in terms of solutions of the forward equation by using the duality relation (3.11). We find that for any $\psi \in M[0, 1]$,

$$n_\varepsilon(t, \cdot\,; \psi) \to n(t, \cdot\,; \psi) \quad \text{as } \varepsilon \downarrow 0$$

where convergence holds with respect to the weak* topology of $X^* = M[0, 1]$, and is uniform for t in bounded intervals of $(0, \infty)$.

4. Discussion. In the previous two sections we have proved the essential correctness of two limit arguments initially derived in a heuristic manner. We expect these cases to be exemplary for a general procedure: (i) start imagining how any model simplification works on the level of the individual, (ii) take good care that birth rates keep behaving, (iii) translate individual behavior into a structured population model, both before and after the simplification, (iv) use the Trotter–Kato theorem to connect the two. The upshot from the examples discussed in this paper is that our intuition derived from the individual level appears to be essentially correct when applied to the population level, at least when we are careful. To emphasize the latter point we finish with three cautionary notes.

(i) From a biological point of view the models from which we started in our examples were already fairly metaphorical. In deriving them we made a great number of simplifying assumptions about the underlying biology, comparable to the ones we spelled out in our limit arguments. The nice thing about apparently being able to make our simplifications with impunity already at the level of the individual, is that usually for the more complicated pictures of individual behavior that lie at the start of our considerations we do not even know how to formulate a full population model. Yet, it is of great importance not to be too naive about our simplifications. A thorough analysis of some metaphorical examples such as those we consider in this paper should

help to clarify the issues. In this context we may point to the work of Chipot and Edelstein (1983) on the dynamics of fungal pellet cultures. Their heuristic model formulation basically seems comparable to the formulation that we chose in our second example. Therefore we feel that the limit model embodied in (1.10) also should be the correct model formulation for that particular class of biological systems, and we fail to understand the rationale that led these authors to a different type of equation.

(ii) The Trotter–Kato theorem only gives information about what happens in finite time intervals. Often our main interest is in the long-term behavior of the population model under consideration. Whether the limit argument extends to such properties has to be ascertained in a separate manner. As an example we may refer to Heijmans (1984) who considers both the transient behavior and some properties of the stable i-state distribution (the dominant eigenfunction of the forward equation), as well as the eventual convergence of the p-state towards this distribution, for a model of satiation dependent predatory behavior.

(iii) The proofs in this paper only apply to the linear time-invariant case, i.e., we did not allow any direct or indirect interactions between the individuals. Ultimately, we will wish to extend the limit theorems to the nonlinear case as well. After all, the greatest strength of the structured population methodology is that it allows us for the first time to incorporate various biologically realistic mechanisms for density dependent population regulation, such as a feedback through the limiting of individual growth by food shortage, into analytically formulated population models. Two approaches are possible. Either we could take recourse to direct nonlinear extensions of the Trotter–Kato theorem (compare, e.g., Clément, Heijmans et al. (1987, § 2.3)), or we could try to fall back on the specific mathematical structure of the equations of structured populations, whose main property is that for a given course of the environment the equations are linear (but time-inhomogeneous). Abstractly, such equations take the form

$$\frac{dn}{dt}(t) = A_\varepsilon(E(t))n(t), \qquad n(0) = \phi \in X. \tag{4.1}$$

Here the vector $E(t)$ describes the environment at time t, and can be calculated as the p-output

$$E(t) = \langle n(t), \xi \rangle$$

for some $\xi \in X^*$. Assuming that, for a given input $E(\cdot)$, the linear time-inhomogeneous equation (4.1) has a solution $n(\cdot)$ we can compute the p-output

$$\tilde{E}(t) = \langle n(t), \xi \rangle.$$

Solving (4.1) amounts to solving the fixed-point equation

$$\tilde{E}(\cdot) = E(\cdot).$$

This fixed-point equation still depends on the parameter ε. If this dependence is continuous (in a sense to be specified) then we might expect that the same is true for its solution.

However, all this is music of a distant future as only the first hesitant steps toward a proof of existence and uniqueness theorems for somewhat more general structured population models of the form (4.1) are being taken at this very moment. Therefore the present paper should only be considered as an introduction to the fascinating problem of putting a more rigorous basis under the structured population methodology.

REFERENCES

P. L. BUTZER AND H. BERENS (1967), *Semigroups of Operators and Approximation*, Springer-Verlag, Berlin, New York.

M. CHIPOT AND L. EDELSTEIN (1983), *A mathematical theory of size distributions in tissue culture*, J. Math. Biol., 16, pp. 115–130.

PH. CLÉMENT, O. DIEKMANN, M. GYLLENBERG, H. J. A. M. HEIJMANS, AND H. R. THIEME (1987), *Perturbation theory for dual semigroups, I. The sun-reflexive case*, Math. Ann., 277, pp. 709–725.

PH. CLÉMENT, H. J. A. M. HEIJMANS ET AL. (1987). *One-Parameter Semigroups*, CWI Monographs 5, North-Holland, Amsterdam.

O. DIEKMANN, H. J. A. M. HEIJMANS, AND H. R. THIEME (1984), *On the stability of the cell size distribution*, J. Math. Biol., 19, pp. 227–248.

L. EDELSTEIN AND Y. HADAR (1983), *A model for pellet size distributions in submerged mycelial cultures*, J. Theoret. Biol., 105, pp. 427–452.

H. J. A. M. HEIJMANS (1984), *Holling's 'hungry mantid' model for the invertebrate functional response considered as a Markov process, Part* III: *Stable satiation distribution*, J. Math. Biol., 21, pp. 115–143.

——— (1986), *Structured populations, linear semigroups, and positivity*, Math. Z., 191, pp. 599–617.

J. A. J. METZ AND O. DIEKMANN (1986), *The Dynamics of Physiologically Structured Populations*, Lecture Notes in Biomathematics 68, Springer-Verlag, Berlin.

J. A. J. METZ, A. M. DE ROOS, AND F. VAN DEN BOSCH (1988), *Population Models Incorporating Physiological Structure*, in Size Structured Populations; Ecology and Evolution, L. Persson and B. Ebenman, eds., Springer-Verlag, Berlin, New York, pp. 106–126.

L. F. MURPHY (1983), *A nonlinear growth mechanism in size structured population dynamics*, J. Theoret. Biol., 104, pp. 493–506.

R. NAGEL, ED. (1986), *One-Parameter Semigroups of Positive Operators*, Lecture Notes in Mathematics, Springer-Verlag, Berlin.

A. PAZY (1983), *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York.

J. W. SINKO AND W. STRIEFER (1967), *A new model for age-size structure of a population*, Ecology, 48, pp. 910–918.

W. STREIFER (1974), *Realistic models in population ecology*, in Advances in Ecological Research, 8, A. MacFayden, ed., pp. 199–266.

H. F. TROTTER (1958), *Approximation of semigroups of operators*, Pacific J. Math., 8, pp. 887–919.

J. VAN SICKLE (1977), *Analysis of a distributed parameter population model based on physiological age*, J. Theoret. Biol., 64, pp. 571–586.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 886-907, July 1989

007

SOME BLOWUP RESULTS FOR A NONLINEAR PARABOLIC EQUATION WITH A GRADIENT TERM*

M. CHIPOT† AND F. B. WEISSLER‡

Abstract. Under some conditions, a blowup result is proved for the solution u of:

$$u_t = \Delta u - |\nabla u|^q + |u|^{p-1}u, \quad t>0, \quad x\in\Omega,$$
$$u(t,x) = 0, \quad t>0, \quad x\in\Gamma,$$
$$u(0,x) = \varphi(x), \quad x\in\Omega.$$

The associated elliptic problem is also studied.

Key words. blowup, nonlinear parabolic equations

AMS(MOS) subject classifications. 35K60, 35B35, 35B60

1. Introduction. In this paper we study the solution of the following semilinear parabolic problem:

$$\begin{aligned} &u_t = \Delta u - |\nabla u|^q + |u|^{p-1}u, && t>0, \quad x\in\Omega, \\ &u(t,y) = 0, && t>0, \quad y\in\Gamma, \\ &u(0,x) = \phi(x), && x\in\Omega. \end{aligned} \tag{1.1}$$

Here $\Omega\subset\mathbf{R}^N$ is a bounded domain with smooth boundary Γ, $u=u(t,x)$, Δ and ∇ apply only to the spatial variables, and $p>1$ and $q>1$ are fixed (finite) parameters. Our main goal is to show that under appropriate conditions on q, p, and n, there exists a suitable initial value ϕ so that the corresponding solution of (1.1) blows up in a finite time.

In the case where there is no gradient term, i.e.,

$$\begin{aligned} &u_t = \Delta u + |u|^{p-1}u, && t>0, \quad x\in\Omega, \\ &u(t,y) = 0, && t>0, \quad y\in\Gamma, \\ &u(0,x) = \phi(x), && x\in\Omega, \end{aligned} \tag{1.2}$$

the following result due to Levine [35] has been known for some time (see also Ball [2]).

THEOREM 1.1. *Let $p>1$ and let $\phi:\bar{\Omega}\to R$ be sufficiently smooth (e.g., C^2) with $\phi|_\Gamma\equiv 0$. If ϕ is large enough in the sense that its "energy,"*

$$E(\phi) = \frac{1}{2}\|\nabla\phi\|_2^2 - \frac{1}{p+1}\|\phi\|_{p+1}^{p+1}, \tag{1.3}$$

is negative, then the corresponding solution of (1.2) *blows up in finite time.*

We remark that local existence of solutions for (1.2) follows by standard iteration methods (see, for example, Segal [28]) on the Banach space $C_0(\Omega)$. Thus, if the existence time T of the maximal solution to (1.2) is finite, i.e., if the solution blows up in finite time T, then $\lim_{t\to T}\|u(t,\cdot)\|_\infty = \infty$.

* Received by the editors March 1, 1988; accepted for publication (in revised form) June 21, 1988. This work was done while both authors were at the Institute for Mathematics and Its Applications (IMA), University of Minnesota, Minneapolis, Minnesota 55455, as research fellows during 1984-1985.

† Université de Metz, Département de Mathématiques et Informatique, Ile du Saulcy, 57045 Metz-Cedex, France.

‡ Department of Mathematics, Texas A&M University, College Station, Texas 77843. The work of this author was supported in part by National Science Foundation grant DMS 8201639.

In the past few years, a great deal of work has been done to study the precise behavior of solutions to (1.2) as t approaches the finite blowup time (see [3], [12], [14]-[16], [24], [25], [29], [32], [33]). A corresponding theory is also being developed in the case where $|u|^{p-1}u$ in (1.2) is replaced by λe^u (see [4]-[8], [12], [22], [23], [30]).

We are naturally led to consider more general parabolic problems of the form

$$u_t = \Delta u + f(u, \nabla u). \tag{1.4}$$

To our knowledge, there has not been much study of solutions to equations of the form (1.4) that blow up in finite time. (For an example, see [11].) Moreover, we are not aware of any finite-time blowup results that would apply to (1.1). Furthermore, the gradient term in (1.1) has a damping effect, working against blowup; and so it is not clear if problem (1.1) has solutions that blow up in finite time. Our goal is therefore somewhat modest: to find an analogue of Theorem 1.1 for problem (1.1).

THEOREM 1.2. *Let* $1 < q \leqq 2p/(p+1)$ *and let* $\phi \in W^{3,s}(\Omega)$ *for* s *sufficiently large,* ϕ *not identically zero. In addition suppose the following*:

(i) $\phi = 0$ *on* Γ;
(ii) $\Delta\phi - |\nabla\phi|^q + |\phi|^{p-1}\phi = 0$ *on* Γ;
(iii) $\phi \geqq 0$ *in* Ω;
(iv) $\Delta\phi - |\nabla\phi|^q + \phi^p \geqq 0$ *in* Ω;
(v) $E(\phi) \leqq 0$;
(vi) *If* $q < 2p/(p+1)$, *then* $\|\phi\|_{p+1}$ *is sufficiently large*;
(vii) *If* $q = 2p/(p+1)$, *then* p *is sufficiently large.*

Then the corresponding solution of (1.1) *blows up in finite time, in the* L^∞ *norm.*

The obvious difficulty with this result is that it is not at all clear if such a ϕ exists. A natural candidate for ϕ is a regular solution of the following elliptic problem:

$$\begin{aligned} &\Delta\phi - |\nabla\phi|^q + \lambda\phi^p = 0 \quad \text{in } \Omega, \\ &\phi > 0 \quad \text{in } \Omega, \\ &\phi = 0 \quad \text{on } \Gamma, \end{aligned} \tag{1.5}$$

where $\lambda > 0$ is sufficiently small.

THEOREM 1.3. *Let* $\Omega = B_R = \{x \in \mathbf{R}^n : |x| < R\}$. *Suppose* $1 < q < 2p/(p+1)$ *and (if* $n \geqq 3$), $p < (n+2)/(n-2)$. *Then for all* $\lambda > 0$ *there exists a regular solution* ϕ *of* (1.5). *If* λ *is sufficiently small, then* ϕ *satisfies conditions* (i)-(vi) *in Theorem* 1.2.

Suppose $n = 1$ *and* $q = 2p/(p+1)$. *Then for all* $\lambda > \lambda_p$, *where*

$$\lambda_p = \frac{(2p)^p}{(p+1)^{2p+1}}, \tag{1.6}$$

there exists a regular solution ϕ *of* (1.5). *If in addition* $\lambda \leqq 2/(p+1)$, *then* ϕ *satisfies conditions* (i)-(v) *in Theorem* 1.2. ("*Regular*" *above means regular enough to apply Theorem* 1.2.)

The paper is organized as follows. In § 2 we prove local existence and uniqueness, and regularity for problem (1.1) with initial values in an appropriate Sobolev space. Moreover, we indicate precisely the conditions on s required for Theorem 1.2 and prove that conditions (i)-(iv) on ϕ imply $u(t, \cdot) \geqq 0$ and $u_t(t, \cdot) \geqq 0$ throughout the trajectory. In § 3 we prove Theorem 1.2, using energy arguments based on the methods found in Ball [2]. We have attempted to write § 3 so that it is, at least formally, independent of the technicalities of § 2. In § 4 we begin the study of (1.5) and prove Theorem 1.3. Finally, in § 5 we present some additional results concerning (1.5). In particular, we show that the value of λ_p claimed in Theorem 1.3 is in fact sharp.

We remark that the value $q=2p/(p+1)$ is "critical" in many respects. The condition $q \leqq 2p/(p+1)$ arises naturally in the energy arguments. When $q=2p/(p+1)$, both (1.1) and (1.5) have the same scaling properties as the same equations without the gradient term; and the character of solutions to (1.5) changes considerably as q is smaller than, equal to, or bigger than $2p/(p+1)$.

2. Local existence and regularity for the evolution equation. In this section $\Omega \subset \mathbf{R}^n$ is a bounded domain with smooth boundary $\Gamma = \partial\Omega$. Also, p and q are fixed real numbers, strictly larger than 1. Our goal is to construct a local theory for the parabolic problem (1.1). The first step is to write the corresponding variation of parameters integral equation

$$u(t) = e^{t\Delta}\phi + \int_0^t e^{(t-s)\Delta} J(u(s))\, ds, \tag{2.1}$$

where $u(t) = u(t, \cdot)$ and $J = J_1 + J_2$ with

$$J_1(u) = -|\nabla u|^q, \qquad J_2(u) = |u|^{p-1}u.$$

Also, $e^{t\Delta}$ denotes the heat semigroup on Ω with Dirichlet boundary conditions. Recall that for $1 < s < \infty$, $e^{t\Delta}$ is an analytic, contraction, C_0 semigroup on $L^s = L^s(\Omega)$. Furthermore, the domain of its generator in L^s is

$$D_s(\Delta) = W^{2,s}(\Omega) \cap W_0^{1,s}(\Omega).$$

Moreover, it is known [10] that $e^{t\Delta}$ restricts to a C_0 semigroup on $W_0^{1,s}(\Omega)$, $1 < s < \infty$.

We will construct a local theory for the integral equation (2.1) in the Banach space $W_0^{1,s} = W_0^{1,s}(\Omega)$, s sufficiently large, using the framework developed in [31]. Note that for suitable $r_1, r_2 \geqq 1$,

$$J_1 : W_0^{1,s} \to L^{r_1}, \qquad J_2 : W_0^{1,s} \to L^{r_2}$$

are continuously Fréchet differentiable maps, Lipschitz on bounded sets in $W_0^{1,s}$. Indeed, r_1 can clearly be chosen s/q, provided $s \geqq q$, and allowable values for r_2 can easily be computed by first determining when $W_0^{1,s}$ is embedded in $L^{r_2 p}$. Moreover, if $1 \leqq r \leqq s < \infty$, then for $t > 0$

$$e^{t\Delta} : L^r \to W_0^{1,s}$$

is bounded with norm bounded by $Ct^{-\alpha}$, where

$$\alpha = \frac{n}{2}\left(\frac{1}{r} - \frac{1}{s}\right) + \frac{1}{2}$$

and C can be chosen uniformly up to any finite time. (See [31, Lemma 4.1] and [1, Thm. 4.17].) Therefore, for each $t > 0$, the map $K_t \equiv e^{t\Delta}J$ is a continuously Fréchet differentiable mapping of $W_0^{1,s}$ into itself, Lipschitz on bounded sets. To apply the results in [31], it suffices to choose s so that $\alpha < 1$ with both $r = r_1 \geqq 1$ and $r = r_2 \geqq 1$. Routine calculations (albeit somewhat tedious) show that this can be done if

$$\begin{aligned} &s \geqq q, \qquad s > n(q-1), \\ &s \geqq np/(n+p), \qquad s > n(p-1)/(p+1). \end{aligned} \tag{2.2}$$

(The conditions on the left side of (2.2) come from the requirement that $r_1, r_2 \geqq 1$; the conditions on the right side come from the requirement that $\alpha < 1$ for both r_1 and r_2.) From now on we assume that s satisfies (2.2). (Later we will need an additional assumption on s.) Thus, by Theorem 1 of [31], for every $\phi \in W_0^{1,s}$, there is a unique

maximal solution $u \in C([0, T_\phi); W_0^{1,s})$ to the integral equation (2.1). T_ϕ is the existence time of the solution starting at ϕ; and if $T_\phi < \infty$, then $\|u(t)\|_{W_0^{1,s}} \to \infty$ as $t \to T_\phi$.

We remark that if $q < 2$, then a local theory for the integral equation (2.1) can be constructed in $L^r(\Omega)$, using the framework developed in [34]. In fact, Theorem 2 of [34] needs to be modified slightly to handle a nonlinearity of the form $J = J_1 + J_2$. (Two spaces E_{J_1} and E_{J_2} are needed instead of just E_J.) We omit the details and simply indicate that if

$$\begin{aligned} &r > n(p-1)/2, \\ &r > n(q-1)/(2-q), \qquad q < 2, \end{aligned} \tag{2.2a}$$

then we have local existence and uniqueness for (2.1) in L^r. In particular, if the existence time T_ϕ is finite, then $\|u(t, \cdot)\|_r \to \infty$ as $t \to T_\phi$. It follows, of course, that $\|u(t, \cdot)\|_\infty \to \infty$ as $t \to T_\phi$. Since $q \leqq 2p/(p+1)$ implies $q < 2$, this is the case under the hypotheses of Theorem 1.2.

We would like to show that if ϕ is sufficiently regular, then the resulting solution of (2.1) is also a solution of the original problem (1.1) and has some additional regularity properties. Recall that the integral equation (2.1) gives rise to a "semiflow" W_t on $W_0^{1,s}$, i.e., W_t takes $\phi \in W_0^{1,s}$ to its value at time t under the action of (2.1). In other words, $W_t\phi = u(t)$, where $u(t)$ is the maximal solution in $W_0^{1,s}$ of (2.1) with initial value ϕ. In particular, $W_t\phi$ is defined precisely for $t \in [0, T_\phi)$. The generator of the semiflow W_t is

$$B\phi = \lim_{t \to 0^+} \frac{W_t\phi - \phi}{t}, \tag{2.3}$$

where the limit is taken in $W_0^{1,s}$. The domain of B, i.e., $D(B)$, is simply the set of $\phi \in W_0^{1,s}$ for which the limit (2.3) exists. Formally, $B = \Delta + J$, i.e.,

$$B\phi = \Delta\phi - |\nabla\phi|^q + |\phi|^{p-1}\phi. \tag{2.4}$$

However, the characterization of B and $D(B)$ in Theorem 3.1 of [31] is somewhat abstract, and some care is needed to describe B and $D(B)$ in the present context. For technical convenience, we make the following additional restrictions on s:

$$s \geqq 2q, \qquad s > nq. \tag{2.5}$$

PROPOSITION 2.1. *Suppose $s \in \mathbf{R}$ satisfies* (2.2) *and* (2.5). *Let W_t be the semiflow on $W_0^{1,s}$ resulting from the integral equation* (2.1), *with generator B and domain $D(B)$. Then $D(B)$ is the set of all $\phi \in W^{3,s} \cap W_0^{1,s}$ such that*

$$\Delta\phi - |\nabla\phi|^q + |\phi|^{p-1}\phi \in W_0^{1,s}. \tag{2.6}$$

For $\phi \in D(B)$, $B\phi$ is given by (2.4).

Proof. First suppose $\phi \in D(B)$. By Theorem 3.1(iv) in [31], it follows that

$$\Delta e^{t\Delta}\phi + e^{t\Delta}(-|\nabla\phi|^q + |\phi|^{p-1}\phi) \tag{2.7}$$

converges strongly in $W_0^{1,s}$ to $B\phi$ as $t \to 0^+$. Now certainly $|\nabla\phi|^q \in L^{s/q}$ and, thanks to (2.5), $|\phi|^{p-1}\phi \in W_0^{1,s} \subset L^{s/q}$. Thus, the second term in (2.7) converges in $L^{s/q}$ as $t \to 0^+$ to $(-|\nabla\phi|^q + |\phi|^{p-1}\phi)$. Furthermore, again by (2.5), $\phi \in W_0^{1,s} \subset H_0^1(\Omega)$; and so $\Delta e^{t\Delta}\phi \to \Delta\phi$ in $H^{-1}(\Omega)$ as $t \to 0^+$. Thus, the distributional limit of (2.7) as $t \to 0^+$ is the desired expression

$$\Delta\phi - |\nabla\phi|^q + |\phi|^{p-1}\phi.$$

This must be the same as the $W_0^{1,s}$ limit, which proves (2.6) and (2.4).

To prove that $\phi \in W^{3,s}$, note first that $B\phi \in W_0^{1,s} \subset L^{s/q}$ and $-|\nabla\phi|^q + |\phi|^{p-1}\phi \in L^{s/q}$. Hence $\Delta\phi \in L^{s/q}$. Since $\phi \in W_0^{1,s/q}$, elliptic regularity gives us that $\phi \in W^{2,s/q}$. Therefore $\nabla\phi \in W^{1,s/q}$. Again by (2.5), $W^{1,s} \subset W^{1,s/q} \subset L^\infty$. Thus $|\nabla\phi|^q$ and $|\phi|^{p-1}\phi$ are both in L^∞. Since $B\phi \in W_0^{1,s} \subset L^\infty$, it follows that $\Delta\phi \in L^\infty$. Thus $\phi \in W^{2,r}(\Omega)$ for every finite r. Therefore $|\nabla\phi|^q$ and $|\phi|^{p-1}\phi$ are both in $W^{1,s}$; and since $B\phi \in W^{1,s}$, we get that $\Delta\phi \in W^{1,s}$. By higher-order elliptic regularity (see, for example, [9, Thm. IX.32]), it follows that $\phi \in W^{3,s}$.

On the other hand, suppose $\phi \in W^{3,s} \cap W_0^{1,s}$ satisfies (2.6). To show that $\phi \in D(B)$, we must show, by Theorem 3.1(iv) in [31], that for all $t > 0$, $e^{t\Delta}\phi$ is in the domain of Δ as a semigroup generator in $W_0^{1,s}$ and that (2.7) has a limit in $W_0^{1,s}$ as $t \to 0^+$. Now $e^{t\Delta}$ is an analytic semigroup on L^s. Thus, $e^{t\Delta}$ restricts us to an analytic semigroup on $D_s(\Delta)$, considered as a Banach space with its graph norm. Since $\phi \in D_s(\Delta) = W^{2,s} \cap W_0^{1,s}$, it follows that for all $t > 0$, $e^{t\Delta}\phi$ is in the domain of Δ as a semigroup generator in $D_s(\Delta)$. Since $D_s(\Delta)$ is continuously embedded in $W_0^{1,s}$, $e^{t\Delta}\phi$ is in the domain of Δ as a semigroup generator in $W_0^{1,s}$. Finally, again since $\phi \in D_s(\Delta)$, $\Delta e^{t\Delta}\phi = e^{t\Delta}\Delta\phi$, where both expressions make sense in L^s. Consequently, (2.7) equals

$$e^{t\Delta}(\Delta\phi - |\nabla\phi|^q + |\phi|^{p-1}\phi),$$

which clearly has a limit in $W_0^{1,s}$ as $t \to 0^+$ because of (2.6).

Remark. The proof above shows that Proposition 2.1 remains correct if $W^{3,s}$ is replaced by $W^{2,s}$. Higher-order elliptic regularity allowed us to conclude $\phi \in W^{3,s}$.

PROPOSITION 2.2. *Under the same conditions as in Proposition* 2.1, *let* $\phi \in D(B)$ *and* $u(t) = W_t\phi$; *i.e.*, $u(t)$ *is the maximal solution of* (2.1). *Then we have the following*:

(i) $u \in C^1([0, T_\phi); W_0^{1,s})$ *and*

$$u'(t) = \Delta u(t) - |\nabla u(t)|^q + |u(t)|^{p-1}u(t), \tag{2.8}$$

where each term on the right side of (2.8) *is in* $C([0, T_\phi); L^{s/q})$;

(ii) $u \in C([0, T_\phi); W^{2,s/q})$;

(iii) $\|u(t)\|_\infty$ *and* $\|\nabla u(t)\|_\infty$ *are bounded on any interval* $[0, T]$ *with* $T < T_\phi$.

Proof. By Theorem 2.2 in [31], $u \in C^1([0, T_\phi); W_0^{1,s})$, $u(t) \in D(B)$ for all $t \in [0, T_\phi)$, and $u'(t) = Bu(t)$. The previous proposition now implies (2.8). Furthermore, $|\nabla u(t)|^q$ and $|u(t)|^{p-1}u(t)$ are clearly continuous into $L^{s/q}$ (again using (2.5)), and since $Bu(t)$ is continuous into $W_0^{1,s}$, it follows that $\Delta u(t)$ is continuous into $L^{s/q}$. This proves (i).

Since $u(t) \in D(B)$, the previous proposition implies $u(t) \in W^{2,s/q} \cap W_0^{1,s/q}$. Also $u(t)$ and $\Delta u(t)$ are both continuous in $L^{s/q}$. Since the graph norm for Δ on $W^{2,s/q} \cap W_0^{1,s/q}$ is equivalent to the $W^{2,s/q}$ norm, it follows that $u(t)$ is continuous into $W^{2,s/q}$, which proves (ii).

Finally, (iii) follows easily since $W^{1,s/q}$ is continuously embedded in L^∞, thanks to assumption (2.5).

PROPOSITION 2.3. *Under the same conditions as in Proposition* 2.1, *let* $\phi \in W_0^{1,s}$ *with* $\phi \geqq 0$ *almost everywhere in* Ω. *Then* $u(t) = W_t\phi \geqq 0$ *for all* $t \in [0, T_\phi)$.

Proof. Any $\phi \geqq 0$ in $W_0^{1,s}$ can be approximated in $W_0^{1,s}$ by nonnegative C^∞ functions with compact support in Ω, in particular by nonnegative functions in $D(B)$. By the continuity properties of the semiflow W_t (see [31, Thm. 1]), we may therefore assume $\phi \geqq 0$ is in $D(B)$. (In fact, we only need the result for $\phi \in D(B)$.)

Multiplying (2.8) by

$$u(t)^- \equiv \frac{|u(t)| - u(t)}{2}$$

and integrating over Ω yields

$$\int u'u^- = \int (\Delta u)u^- - \int |\nabla u|^q u^- + \int |u|^{p-1}uu^-,$$

where we have suppressed the dependence on t. By the previous proposition, we clearly have $u \in C^1((0, T_\phi); L^2)$ and $u \in C((0, T_\phi); H_0^1)$. Thus

$$\int u'u^- = -\frac{1}{2}\frac{d}{dt}\int (u^-)^2$$

and

$$\int (\Delta u)u^- = \int |\nabla u^-|^2.$$

The first formula above follows from the definition of u^- (multiply out $(u^-)^2$) and the second formula from well-known facts about ∇u^- [17, § 7.4]. We now restrict ourselves to $t \in [0, T]$ for a fixed $T < T_\phi$. By part (iii) of the previous proposition, there is a constant $C = C(T)$ such that

$$\int |u|^p |u^-| \leqq C \int |u^-|^2$$

and

$$\begin{aligned}\int |\nabla u|^q |u^-| &\leqq C \int |\nabla u^-||u^-| \\ &\leqq \varepsilon C \int |\nabla u^-|^2 + C_\varepsilon C \int |u^-|^2,\end{aligned}$$

where $\varepsilon > 0$ is arbitrary, but C_ε depends on the choice of ε. Putting all this together, and choosing $\varepsilon > 0$ so that $\varepsilon C \leqq 1$, we get that for $t \in (0, T]$,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int |u^-|^2 &\leqq -\int |\nabla u^-|^2 + \varepsilon C \int |\nabla u^-|^2 + C_\varepsilon C \int |u^-|^2 \\ &\leqq C_\varepsilon C \int |u^-|^2.\end{aligned}$$

Since $u^- \in C([0, T]; L^2)$ and $u(0)^- = \phi^- = 0$, Gronwall's lemma now implies $\int |u^-|^2 = 0$ for all $t \in [0, T]$. Since $T < T_\phi$ is arbitrary, we see that $u(t)^- = 0$ for all $t \in [0, T_\phi)$.

For the energy arguments in the next section we need not only $u(t) \geqq 0$, but also $u'(t) \geqq 0$, throughout the trajectory. To prove this with the weak maximum principle methods of previous proof, we need some higher-order regularity in t. We begin with the following lemma. Its proof is modeled on the proof of Theorem 3 in [20]. (See also [31, Prop. 1.2].)

LEMMA 2.4. *Under the same conditions as in Proposition* 2.1, *let* $\phi \in D(B)$ *and* $u(t) = W_t\phi \in C^1([0, T_\phi); W_0^{1,s})$. *Denote* $v(t) = u'(t)$, *so* $v \in C([0, T_\phi); W_0^{1,s})$. *Then for any compact subinterval* $[\varepsilon, T] \subset (0, T_\phi)$, $v : [\varepsilon, T] \to W_0^{1,s}$ *is Hölder continuous.*

Proof. By Theorem 2.2 of [31], $v(t)$ satisfies the following integral equation:

$$\begin{aligned}v(t) = e^{t\Delta}(B\phi) - q\int_0^t e^{(t-s)\Delta}(|\nabla u(s)|^{q-2}\nabla u(s)\cdot\nabla v(s))\,ds \\ + p\int_0^t e^{(t-s)\Delta}|u(s)|^{p-1}v(s)\,ds.\end{aligned} \tag{2.9}$$

Since $B\phi \in W_0^{1,s} \subset L^s$ and $e^{t\Delta}$ is an analytic semigroup on L^s, it follows that $e^{t\Delta}(B\phi)$ is in $C^1((0,\infty); D_s(\Delta))$ and thus is certainly Hölder continuous in $W_0^{1,s}$ on $[\varepsilon, T]$. Thus, it suffices to show that the two integral terms are Hölder continuous on $[\varepsilon, T]$. We consider only the first one, the second one being easier to handle.

By Proposition 2.2(ii) and the fact that $W^{1,s/q} \subset L^\infty$ by (2.5), we have that ∇u, and hence $|\nabla u|^{q-2}\nabla u$ must be in $C([0, T]; L^\infty)$. (Obviously, we mean that $|\nabla u|^{q-2}\nabla u = 0$ in the case where $|\nabla u| = 0$. This presents no problem since $q > 1$.) Moreover, ∇v is clearly in $C([0, T]; L^s)$. Therefore

$$w(t) \equiv |\nabla u(t)|^{q-2}\nabla u(t) \cdot \nabla v(t) \tag{2.10}$$

is in $C([0, T]; L^s)$. Let

$$z(t) = \int_0^t e^{(t-s)\Delta} w(s)\, ds.$$

Since $W_0^{1,s} = D_s(\sqrt{-\Delta})$, the domain of $\sqrt{-\Delta}$ in L^s (see [26], [27]), to show that $z(t)$ is Hölder continuous in $W_0^{1,s}$ it suffices to show $\sqrt{-\Delta}\, z(t)$ Hölder continuous in L^s. For $0 \leqq t < t+\tau \leqq T$, we have

$$\begin{aligned}
\sqrt{-\Delta}\,(z(t+\tau)-z(t)) &= \sqrt{-\Delta}\,(e^{\tau\Delta}-I)\int_0^t e^{(t-s)\Delta} w(s)\, ds \\
&\quad + \sqrt{-\Delta}\int_0^\tau e^{s\Delta} w(t+\tau-s)\, ds \\
&= (e^{\tau\Delta}-I)(-\Delta)^{-\alpha}\int (-\Delta)^{\alpha+1/2} e^{(t-s)\Delta} w(s)\, ds \\
&\quad + \int_0^\tau (-\Delta)^{1/2} e^{s\Delta} w(t+\tau-s)\, ds,
\end{aligned}$$

where $0 < \alpha < \frac{1}{2}$. Using the facts [21, Thms. 11.3, 12.1] that for $t \in (0, T]$ and $0 < \nu < 1$

$$\|(-\Delta)^\nu e^{t\Delta}\|_s \leqq Ct^{-\nu}, \qquad \|(e^{t\Delta}-I)(-\Delta)^{-\nu}\|_s \leqq Ct^{\nu},$$

we deduce that

$$\begin{aligned}
\|\sqrt{-\Delta}\,(z(t+\tau)-z(t))\|_s &\leqq C\tau^\alpha \int_0^t (t-s)^{-\alpha-1/2}\, ds \sup_{[0,T]} \|w(t)\|_s \\
&\quad + C\int_0^\tau s^{-1/2}\, ds \sup_{[0,T]} \|w(t)\|_s.
\end{aligned}$$

This proves the Hölder continuity of $\sqrt{-\Delta}\, z(t)$ in L^s and thereby completes the proof of the lemma.

Remark. For the above result we do not need the rather strong result that $W_0^{1,s} = D_s(\sqrt{-\Delta})$. The easier result that $D_s((-\Delta)^\nu)$ is continuously embedded in $W_0^{1,s}$ [21, Thm. 9.2] can be used with only a slight modification of the proof.

PROPOSITION 2.5. *Under the same conditions and with the same notation of Lemma* 2.4, *we have that* $v \in C^1((0, T_\phi); L^{s/q})$ *and*

$$v'(t) = \Delta v(t) - q|\nabla u(t)|^{q-2}\nabla u(t) \cdot \nabla v(t) + p|u(t)|^{p-1} v(t). \tag{2.11}$$

Proof. Since $\phi \in D(B)$, $u \in C^1[(0, T_\phi); W_0^{1,s})$, and so $\nabla u \in C^1([0, T_\phi); L^s)$. Furthermore, by the previous lemma ∇v is Hölder continuous in L^s on $[\varepsilon, T]$. It follows that $w(t)$ given by (2.10) is Hölder continuous in $L^{s/q}$ on $[\varepsilon, T]$. Similarly, $|u|^{p-1}v$ is Hölder continuous in $L^{s/q}$ on $[\varepsilon, T]$.

We now consider the integral equation (2.9) as an equation in $L^{s/q}$. The semigroup is analytic in $L^{s/q}$ and the two integrands (not including $e^{(t-s)\Delta}$) are both Hölder continuous functions into $L^{s/q}$ on any compact subinterval $[\varepsilon, T] \subset (0, T_\phi)$. The result now follows from well-known properties of analytic semigroups. (See, for example, [19, Thm. 1.27, Chap. IX]).

We are now able to prove the desired positivity of $u'(t)$.

PROPOSITION 2.6. *Suppose $s \in \mathbf{R}$ satisfies (2.2) and (2.5) and that ϕ is in $D(B)$. Let $u(t) = W_t\phi$ and $v(t) = u'(t)$, and suppose further that $v(0) = u'(0) = B\phi \geqq 0$. Then $u'(t) \geqq 0$ for all $t \in [0, T_\phi)$.*

Proof. We know that $v \in C([0, T_\phi); W_0^{1,s}) \subset C([0, T_\phi); H_0^1)$ and that $v \in C^1((0, T_\phi); L^{s/q}) \subset C^1((0, T_\phi); L^2)$. Thus, multiplying (2.11) by $v(t)^-$ and integrating over Ω yields

$$\frac{1}{2}\frac{d}{dt}\int |v^-|^2 = -\int |\nabla v^-|^2 + q\int |\nabla u|^{q-2}\nabla u \cdot \nabla v(v^-) - p\int |u|^{p-1}vv^-$$

$$\leqq -\int |\nabla v^-|^2 + C\int |\nabla v^-||v^-| + C\int |v^-|^2,$$

where we use Proposition 2.2(iii) to estimate ∇u and u, and C can be chosen uniformly for $t \in (0, T]$, $T < T_\phi$. The proof is completed exactly as in the proof of Proposition 2.3.

3. Energy arguments. In this section we prove Theorem 1.2. Throughout this section, Ω, Γ, p, and q are as in the Introduction, and we assume $s \in \mathbf{R}$ satisfies (2.2) and (2.5). Also, we take $\phi \in W^{3,s}$, not identically zero, satisfying hypotheses (i)-(iv) of Theorem 1.2. In the language of § 2, that means $\phi \in D(B) \subset W_0^{1,s}$ with $\phi \geqq 0$ and $B\phi \geqq 0$. T_ϕ is the existence time of the maximal solution $u(t)$ of the integral equation (2.1). By Propositions 2.2, 2.3, and 2.6, $u \in C^1([0, T_\phi); W_0^{1,s})$, satisfying equation (2.8), with $u(t) \geqq 0$ and $u'(t) \geqq 0$ for all $t \in [0, T_\phi)$.

LEMMA 3.1. *The energy of the solution $u(t)$,*

$$E(u(t)) = \frac{1}{2}\|\nabla u(t)\|_2^2 - \frac{1}{p+1}\|u(t)\|_{p+1}^{p+1},$$

is a nonincreasing function of $t \in [0, T_\phi)$.

Proof. Since $u \in C^1([0, T_\phi); W_0^{1,s})$ and, by (2.5), $W_0^{1,s} \subset H_0^1$ and $W_0^{1,s} \subset L^\infty$, it follows that $E(u(t))$ is a C^1 function of $t \in [0, T_\phi)$. We easily calculate from (2.8) that

$$\frac{d}{dt}E(u(t)) = \langle -\Delta u(t), u'(t)\rangle - \langle u(t)^p, u'(t)\rangle$$

$$= -\langle u'(t) + |\nabla u(t)|^q, u'(t)\rangle$$

$$\leqq 0.$$

LEMMA 3.2. *Suppose $E(\phi) = E(u(0)) \leqq 0$ and that $q \leqq 2p/(p+1)$. Then for all $t \in [0, T_\phi)$*

$$|\langle u(t), |\nabla u(t)|^q\rangle| \leqq \left(\frac{2}{p+1}\right)^{q/2} C(p, q)\|u(t)\|_{p+1}^{p+1-\alpha}$$

where

$$\alpha = p - \frac{q(p+1)}{2} \geqq 0$$

and $C(p, q) = 1$ in case $q = 2p/(p+1)$.

Proof. From Hölder's inequality, it follows that

$$|\langle u, |\nabla u|^q\rangle| \leqq \|u\|_{p+1} \| |\nabla u|^q \|_{(p+1)/p}$$
$$= \|u\|_{p+1} \|\nabla u\|^q_{q(p+1)/p}.$$

Since $q \leqq 2p/(p+1)$, we have $q(p+1)/p \leqq 2$; and so

$$|\langle u, |\nabla u|^q\rangle| \leqq C(p,q)\|u\|_{p+1}\|\nabla u\|_2^q,$$

where $C(p,q)=1$ if $q=2p/(p+1)$.

By the previous lemma $E(u(t)) \leqq 0$ for all $t \in [0, T_\phi)$, or

$$\|\nabla u\|_2^q \leqq \left(\frac{2}{p+1}\right)^{q/2} \|u\|_{p+1}^{q(p+1)/2}.$$

The result follows by combining the last two inequalities.

Proof of Theorem 1.2. Suppose to the contrary that $T_\phi = \infty$. Let $F(t) = \|u(t)\|_2^2$. Then $F \in C^1([0,\infty))$, $F(0) = \|\phi\|_2^2 > 0$, and

$$F'(t) = 2\langle u(t), u'(t)\rangle$$
$$= -2\|\nabla u(t)\|_2^2 - 2\langle u(t), |\nabla u(t)|^q\rangle + 2\|u(t)\|_{p+1}^{p+1}$$
$$= -4E(t) + 2\left(\frac{p-1}{p+1}\right)\|u(t)\|_{p+1}^{p+1} - 2\langle u, |\nabla u|^q\rangle$$
$$\geqq 2\left(\frac{p-1}{p+1}\right)\|u(t)\|_{p+1}^{p+1} - 2\left(\frac{2}{p+1}\right)^{q/2} C(p,q)\|u(t)\|_{p+1}^{p+1-\alpha},$$

where we have used Lemma 3.1 ($E(u(t)) \leqq E(\phi) \leqq 0$) and Lemma 3.2. Continuing the calculation, we have

$$F'(t) \geqq 2\|u(t)\|_{p+1}^{p+1}\left[\left(\frac{p-1}{p+1}\right) - \left(\frac{2}{p+1}\right)^{q/2} C(p,q)\|u(t)\|_{p+1}^{-\alpha}\right]$$
$$\geqq CF(t)^{(p+1)/2}\left[\left(\frac{p-1}{p+1}\right) - \left(\frac{2}{p+1}\right)^{q/2} C(p,q)\|u(t)\|_{p+1}^{-\alpha}\right].$$

Suppose first $q < 2p/(p+1)$ and that $\|\phi\|_{p+1}$ is sufficiently large so that

$$\left(\frac{p-1}{p+1}\right) - \left(\frac{2}{p+1}\right)^{q/2} C(p,q)\|\phi\|_{p+1}^{-\alpha} \equiv k > 0.$$

Then since $u'(t) \geqq 0$, it follows that

$$F'(t) \geqq kCF(t)^{(p+1)/2}$$

for all $t \in [0,\infty)$. Since $(p+1)/2 > 1$, this is impossible for a function $F \in C^1([0,\infty))$ with $F(0) > 0$. This contradiction shows $T_\phi < \infty$.

Now suppose $q = 2p/(p+1)$. Then $C(p,q) = 1$ and $\alpha = 0$, so

$$F'(t) \geqq CF(t)^{(p+1)/2}\left[\left(\frac{p-1}{p+1}\right) - \left(\frac{2}{p+1}\right)^{p/(p+1)}\right].$$

This again yields a contradiction if p is large enough that the coefficient above is positive.

Remark. The expression

$$\left(\frac{p-1}{p+1}\right) - \left(\frac{2}{p+1}\right)^{p/(p+1)}$$

is increasing in p for $1<p<\infty$ with limit 1 as $p\to\infty$. Thus, if we let p_* be its unique zero in this range, then the above argument works for all $p>p_*$. An easy computation shows $3.3<p_*<3.4$.

4. The elliptic problem. For the moment, Ω, Γ, p, and q are as in the Introduction. Also, λ is always positive. Our goal is first to show the connection between the elliptic problem (1.5) and the hypotheses of Theorem 1.2. Then for $\Omega=B_R$ we will study the existence of solutions to (1.5). Together, this will prove Theorem 1.3.

PROPOSITION 4.1. *Let $2\leqq s<\infty$ and suppose $\phi\in W^{2,s}(\Omega)$ is a solution of* (1.5) *with $\lambda\leqq 2/(p+1)$. Then ϕ satisfies hypotheses* (i)-(v) *of Theorem* 1.2. *If in addition s satisfies* (2.5), *then $\phi\in W^{3,s}$.*

Proof. Hypotheses (i) and (iii) are stated in (1.5), and so there is nothing to prove. Now

$$\Delta\phi-|\nabla\phi|^q+\phi^p=(1-\lambda)\phi^p,$$

which immediately gives (ii) and (iv). Finally, since $\phi=0$ on Γ implies $\phi\in W_0^{1,s}$,

$$\begin{aligned}E(\phi)&=\frac{1}{2}\|\Delta\phi\|_2^2-\frac{1}{p+1}\|\phi\|_{p+1}^{p+1}\\&=-\frac{1}{2}\int_\Omega\phi\,\Delta\phi-\frac{1}{p+1}\int_\Omega\phi^{p+1}\\&=-\frac{1}{2}\int_\Omega\phi|\nabla\phi|^q-\left(\frac{1}{p+1}-\frac{\lambda}{2}\right)\int_\Omega\phi^{p+1}.\end{aligned}$$

Thus $E(\phi)\leqq 0$ because $\lambda\leqq 2/(p+1)$. The regularity of ϕ follows exactly as in the proof of Proposition 2.1. Simply note at the start that, thanks to (2.5), $\phi^p\in W_0^{1,s}$ and so

$$B\phi=(1-\lambda)\phi^p\in W_0^{1,s}.$$

PROPOSITION 4.2. *Assume that $p\leqq(n+2)/(n-2)$. (If $n=1$ or 2, this condition is vacuous.) Suppose that $\phi_k\in H_0^1$, $k=1,2,3,\cdots$, satisfy*

$$\Delta\phi_k+\lambda_k\phi_k^p\geqq 0,\quad \phi_k\geqq 0,\quad \phi_k\not\equiv 0, \tag{4.1}$$

where $\lambda_k>0$ and $\lambda_k\to 0$ as $k\to\infty$. Then $\|\phi_k\|_{p+1}\to\infty$ as $k\to\infty$.

Proof. Suppose not. Then, by passing to a subsequence, we may assume $\|\phi_k\|_{p+1}\leqq M$ independent of k. Let $\|\phi_k\|_{p+1}=N_k$ and

$$\psi_k=\phi_k/N_k.$$

Obviously $\|\psi_k\|_{p+1}=1$. ($N_k\neq 0$ since $\phi_k\not\equiv 0$.) Moreover, multiplying the inequality in (4.1) by ϕ_k and integrating over Ω, we have that

$$\|\nabla\phi_k\|_2^2\leqq\lambda_k\|\phi_k\|_{p+1}^{p+1},$$

or

$$\|\nabla\psi_k\|_2^2\leqq\lambda_k N_k^{p-1}\|\psi_k\|_{p+1}^{p+1}.$$

Since $\|\psi_k\|_{p+1}=1$, $N_k\leqq M$, and $\lambda_k\to 0$, it follows that $\|\nabla\psi_k\|_2\to 0$ as $k\to\infty$, i.e., $\psi_k\to 0$ in $H_0^1(\Omega)$ as $k\to\infty$. However, the condition on p implies that $H_0^1(\Omega)$ is embedded in L^{p+1}, and so $\psi_k\to 0$ in L^{p+1}. This contradicts the fact that $\|\psi_k\|_{p+1}=1$, thereby proving the proposition.

The following corollary to the above two propositions states explicitly how solutions of (1.5) yield solutions of (1.1) that blow up in finite time.

COROLLARY 4.3. *Assume* $s \in \mathbf{R}$ *satisfies* (2.2) *and* (2.5). *Suppose first that* $1 < q < 2p/(p+1)$ *and that* $1 < p < (n+2)/(n-2)$ ($1 < p < \infty$ *if* $n = 1$ *or* 2). *If* $\phi \in W^{2,s}$ *is a solution of* (1.5) *with* λ *sufficiently small, then* ϕ *satisfies all the hypotheses of Theorem* 1.2; *and so the solution of* (1.1) *with initial value* ϕ *blows up in finite time.*

Suppose next that $q = 2p/(p+1)$ *with* $p > p_*$. (*See the remark at the end of* § 3.) *If* $\phi \in W^{2,s}$ *is a solution of* (1.5) *with* $\lambda \leqq 2/(p+1)$, *then* ϕ *satisfies all the hypotheses of Theorem* 1.2; *and so the solution of* (1.1) *with initial value* ϕ *blows up in finite time.*

Now we let $\Omega = B_R = \{x \in \mathbf{R}^n : |x| < R\}$, and we look for solutions of (1.5) on B_R. In fact we are going to look for radially symmetric solutions of (1.5). This is not a genuine restriction, because the techniques of [13] can be used to show that any solution of (1.5) in B_R must be radially symmetric. We are therefore led to consider the following initial value problem:

$$\begin{aligned} &u''(r) + \frac{n-1}{r} u'(r) - |u'(r)|^q + \lambda |u(r)|^p = 0, \qquad r > 0, \\ &u(0) = a > 0, \qquad u'(0) = 0. \end{aligned} \tag{4.2}$$

If $u \in C^2([0, R])$ is a solution of (4.2) with $u(r) > 0$ for $0 \leqq r < R$ and $u(R) = 0$, then $\phi(x) = u(|x|)$ is the desired solution of (1.5). (Note that, for the rest of the paper, we will no longer be directly concerned with problem (1.1). Thus, the letter "u" will henceforth be used to denote solutions of (4.2).)

PROPOSITION 4.4. *Fix* $\lambda > 0$. *For every* $a > 0$ *there exists a* (*unique*) *maximal solution* $u \in C^2([0, R_a))$ *of* (4.2). *Furthermore, we have the following*:

(i) $u'(r) < 0$ *for all* r, $0 < r < R_a$;

(ii) *The function*

$$H(r) = \frac{1}{2} u'(r)^2 + \frac{\lambda}{p+1} |u(r)|^p u(r)$$

is decreasing on $[0, R_a)$;

(iii) *If* $u(r) > 0$ *for all* $0 \leqq r < R_a$, *then* $R_a = \infty$ *and*

$$\lim_{r \to \infty} u(r) = 0, \quad \lim_{r \to \infty} u'(r) = 0, \quad \lim_{r \to \infty} u''(r) = 0.$$

Proof. We first prove the existence of a unique solution to (4.2) on some interval $[0, \varepsilon]$. Consider the system

$$\begin{aligned} u(r) &= a + \int_0^r v(s)\, ds, \\ v(r) &= r^{-(n-1)} \int_0^r s^{n-1} (|v(s)|^q - \lambda |u(s)|^p)\, ds. \end{aligned} \tag{4.3}$$

It is easy to see that a solution of (4.2) is also a solution of (4.3) with $v = u'$. Indeed, simply multiply the equation in (4.2) by r^{n-1} and integrate. On the other hand, by standard iteration techniques, there is certainly a unique solution $u, v \in C([0, \varepsilon])$ to (4.3) for some $\varepsilon > 0$. Clearly, $u \in C^1([0, \varepsilon])$ with $u'(r) = v(r)$. In particular, $u'(0) = 0$. Moreover, v is immediately seen to be in $C^1((0, \varepsilon])$ and so $u \in C^2((0, \varepsilon])$ and satisfies (4.2). It remains to show that u is C^2 at $r = 0$, i.e., that v is C^1 at $r = 0$. From (4.3), l'Hôpital's rule easily gives

$$v'(0) = \lim_{r \to 0} \frac{v(r)}{r} = -\frac{\lambda a^p}{n}.$$

On the other hand, from (4.2)

$$\lim_{r\to 0} v'(r) = \lim_{r\to 0} u''(r) = \frac{(n-1)\lambda a^p}{n} - \lambda a^p = -\frac{\lambda a^p}{n}.$$

Thus, $u \in C^2([0, \varepsilon])$.

Since for $r > 0$, there are no singularities in (4.2), the solution on $[0, \varepsilon]$ can be locally continued to a maximal solution $u \in C^2([0, R_a))$. Since the continuation procedure treats (4.2) as a system $(u(r), v(r))$ with $u'(r) = v(r)$, it follows that if $R_a < \infty$ then either $|u(r)| \to \infty$ or $|u'(r)| \to \infty$ as $r \to R_a$.

To prove (i), note first that $u'(0) = 0$ and $u''(0) < 0$. Hence $u'(r) < 0$ on some interval $(0, \delta)$. Let r_0 be the first positive zero of u'. Then $u(r_0) \neq 0$ since, if $u(r_0) = u'(r_0) = 0$, it follows by uniqueness that $u(r) \equiv 0$. Consequently, from the equation in (4.2)

$$u''(r_0) = -\lambda |u(r_0)|^p < 0,$$

which implies that $u'(r) > 0$ for r in some interval $(r_0 - \varepsilon, r_0)$. This contradicts the choice of r_0, and thereby proves (i).

Next, we compute easily that for $r > 0$

$$\begin{aligned} H'(r) &= u'(r)u''(r) + \lambda |u(r)|^p u'(r) \\ &= u'(r)\left[-\left(\frac{n-1}{r}\right) u'(r) + |u'(r)|^q \right] \\ &< 0. \end{aligned}$$

This proves (ii).

Finally, if $u(r) > 0$ for all $r \in [0, R_a)$, then $0 < H(r) \leqq H(0)$ for all $r \in [0, R_a)$. Thus, $u(r)$ and $u'(r)$ are a priori bounded and so $R_a = \infty$. Now $u'(r) < 0$ and $u(r) > 0$. Hence,

$$\lim_{r\to\infty} u(r) = u_\infty \quad \text{(finite)}$$

exists. Likewise, $H(r)$ has a finite limit as $r \to \infty$. It follows therefore that

$$\lim_{r\to\infty} u'(r) = v_\infty$$

exists. In fact we must have $v_\infty = 0$ for $\lim_{r\to\infty} u(r)$ to exist. Finally, from (4.2) we now deduce that

$$\lim_{r\to\infty} u''(r) = -\lambda |u_\infty|^p.$$

Thus, the only way we can have $\lim_{r\to\infty} u'(r) = 0$ is if $u_\infty = 0$. This completes the proof of (iii).

For a fixed $\lambda > 0$, we denote the first zero of the solution to (4.2) by $z(a)$. We set the convention that $z(a) = \infty$ in case $u(r) > 0$ for all $r \geqq 0$. Thus, the solution $u(r)$ of (4.2) yields the desired solution of (1.5) precisely if $z(a) = R$. This certainly motivates studying the function $z(a)$.

PROPOSITION 4.5. *$\{a > 0: z(a) < \infty\}$ is open and $z(\cdot)$ is continuous on this set. Moreover,*

$$\lim_{a\to 0} z(a) = \infty. \tag{4.4}$$

Also, if $z(a_0) = \infty$ for some $a_0 \in \mathbf{R}$, then $\lim_{a\to a_0} z(a) = \infty$.

Proof. If $z(a) < \infty$, then $u(r) < 0$ for r slightly larger than $z(a)$. By continuous dependence on the data, if we change a by only a little bit, $u(r)$ must still be negative

somewhere, and must therefore have a zero. Continuity of $z(\cdot)$ follows from continuous dependence of $u(r)$ on a and the fact that u can have at most one zero since $u'(r)<0$ for $r>0$.

To prove (4.4), we note that since $H(r)$ is decreasing, it follows that

$$\frac{1}{2}u'(r)^2 \leqq H(r) \leqq H(0) = \frac{\lambda a^{p+1}}{p+1}$$

or

$$|u'(r)| \leqq \sqrt{2\lambda/(p+1)}\, a^{(p+1)/2}. \tag{4.5}$$

Consequently,

$$\begin{aligned} a &= u(0) - u(z(a)) \\ &= -\int_0^{z(a)} u'(s)\, ds \\ &\leqq z(a) a^{(p+1)/2} \sqrt{2\lambda/(p+1)}, \end{aligned}$$

or

$$z(a) \geqq \frac{a^{-(p-1)/2}}{\sqrt{2\lambda/(p+1)}}. \tag{4.6}$$

This proves (4.4).

For the last statement, we show that given $M>0$, then $z(a)>M$ if a is sufficiently close to a_0. Now for $a=a_0$, $u(r)>0$ for all $r>0$; hence $u(r)\geqq\delta>0$ on $[0, M]$. By continuous dependence on the data, if a is sufficiently close to a_0 then $u(r)\geqq\delta/2$ on $[0, M]$. Hence $z(a)>M$ for such a.

Next we would like to study the behavior of $z(a)$ as $a\to\infty$. We first consider the case $q<2p/(p+1)$.

PROPOSITION 4.6. *Assume that* $q<2p/(p+1)$ *and (in the case where* $n\geqq 3$*)* $p<(n+2)/(n-2)$. *Then, for all* $\lambda>0$, *we have*

$$\limsup_{a\to\infty} a^{(p-1)/2} z(a) < \infty, \tag{4.7}$$

$$\lim_{a\to\infty} z(a) = 0. \tag{4.8}$$

Proof. Fix $\lambda>0$. Denote by $u(\cdot\,; a)$ the solution of (4.2) with initial value a; and for all $a>0$, set

$$V_a(r) = a^{-1} u(r a^{-(p-1)/2}; a).$$

Then v_a is easily seen to satisfy

$$\begin{aligned} &v_a'' + \frac{n-1}{r} v_a' - a^{q(p+1)/2-p} |v_a'| + \lambda |v_a|^p = 0, \\ &v_a(0) = 1, \qquad v_a'(0) = 0. \end{aligned} \tag{4.9}$$

Also, $v_a'(r)<0$ for $r>0$ and $v_a(r)>0$ for $0\leqq r< a^{(p-1)/2} z(a)$. Hence

$$0 \leqq v_a(r) \leqq 1, \qquad 0 \leqq r \leqq a^{(p-1)/2} z(a). \tag{4.10}$$

Moreover, (4.5) translates into

$$|v_a'(r)| \leqq \sqrt{2\lambda/(p+1)}, \qquad r \geqq 0. \tag{4.11}$$

Now suppose there exists a sequence $a_m \to \infty$ such that $a_m^{(p-1)/2} z(a_m) \to \infty$ as $m \to \infty$. By the Arzelà-Ascoli theorem and a standard diagonal argument, there is a subsequence, which we still denote a_m, and a continuous function $v:[0,\infty)\to[0,1]$ such that $v_{a_m} \to v$ uniformly on all compact subsets $[0, M] \subset [0,\infty)$. In particular, $v(0)=1$, v is nonincreasing on $[0,\infty)$, and v is Lipschitz continuous with a Lipschitz constant no greater than $\sqrt{2\lambda/(p+1)}$. (Each v_a has these properties.) Finally, since $q < 2p/(p+1)$ and $|v_a'|$ is bounded independent of r and a, it follows from (4.9) that

$$v'' + \frac{n-1}{r} v' + \lambda v^p = 0 \tag{4.12}$$

in the sense of distributions on $(0,\infty)$.

It is well known that since $p < (n+2)/(n-2)$, such a v cannot exist. This is proved on pp. 293-294 of [33] in the case $\lambda = 1$. (See also [18, Prop. 3.9].) The same arguments work for any $\lambda > 0$, or else λ can be scaled away by multiplying v by a suitable factor. This proves (4.7), and hence (4.8).

We now turn to the case $q = 2p/(p+1)$. This will be quite different since the scaled solution v_a satisfies the same equation as u.

LEMMA 4.7. *If $q = 2p/(p+1)$, then for all $a > 0$,*

$$z(a) = a^{-(p-1)/2} z(1). \tag{4.13}$$

Proof. By (4.9), we see that $v_a = u(\cdot\,; 1)$ for all $a > 0$. Hence the first zero of v_a is $z(1)$. However, by the definition of v_a, its first zero is $a^{(p-1)/2} z(a)$. This proves (4.13).

In other words, whether or not $z(a)$ is finite depends entirely on whether or not $z(1)$ is finite. This in turn depends on λ.

LEMMA 4.8. *Let $r_0 \geqq 0$, $\lambda > 0$, $q = 2p/(p+1)$, and suppose $u:(r_0,\infty)\to\mathbf{R}$ is C^2 and satisfies*

(i) $u(r) > 0$, $r > r_0$, *and* $\lim_{r\to\infty} u(r) = 0$;
(ii) $u'(r) < 0$, $r > r_0$, *and* $\lim_{r\to\infty} u'(r) = 0$;
(iii) $u''(r) - |u'(r)|^q + \lambda u(r)^p = 0$, $r > r_0$.

If in addition $u(r)$ satisfies

$$u''(r) \leqq k|u'(r)|^q, \qquad r > r_0, \tag{4.14}$$

where $k > 0$ is a fixed constant, then u must also satisfy

$$u''(r) \leqq \left[1 - \lambda\left(\frac{p+1}{2k}\right)^p\right] |u'(r)|^q, \qquad r > r_0. \tag{4.15}$$

Proof. Since $u'(r) < 0$, inequality (4.14) can be rewritten:

$$(-u'(r))^{-(q-1)} u''(r) \leqq -ku'(r).$$

Integrating this from r to ∞, we get

$$\frac{(-u'(r))^{2-q}}{2-q} \leqq ku(r),$$

or

$$u(r) \geqq \frac{p+1}{2k} (-u'(r))^{2/(p+1)}.$$

(Recall that $q = 2p/(p+1) < 2$.) Hence

$$\begin{aligned} u''(r) &= (-u'(r))^q - \lambda u(r)^p \\ &\leqq \left[1 - \lambda\left(\frac{p+1}{2k}\right)^p\right](-u'(r))^q. \end{aligned}$$

PROPOSITION 4.9. *Suppose* $n=1$ *and* $q=2p/(p+1)$. *If* $\lambda \geqq (2/(p+1))^p$, *then* $z(a)<\infty$ *for all solutions* $u(r)$ *of* (4.2).

Proof. If $z(a)=\infty$, then $u(r)>0$ for all $r>0$. By Proposition 4.4, and the fact that $n=1$, $u(r)$ satisfies conditions (i)-(iii) of Lemma 4.8 with $r_0=0$. Furthermore, by (4.2) with $n=1$, we have that (4.14) holds with $k=1$. Thus (4.15) holds with $k=1$. Since $\lambda \geqq [2/(p+1)]^p$, it follows that $u''(r)\leqq 0$ for $r>0$. This is impossible because $u'(r)<0$ and $u(r)>0$ for $r>0$.

The result in Proposition 4.9 is already enough to give us a solution of (1.5) with $\lambda \leqq 2/(p+1)$. Indeed, $[2/(p+1)]^p < 2/(p+1)$. However, this result can be improved.

LEMMA 4.10. *Suppose* $\lambda > \lambda_p$, *given by* (1.6). *Define the following sequence inductively*:

$$k_0=1, \qquad k_m = 1-\lambda\left(\frac{p+1}{2k_{m-1}}\right)^p,$$

as long as $k_{m-1}>0$. *Then either* k_m *is eventually nonpositive or* $\lim_{m\to\infty} k_m = 0$.

Proof. It is easy to verify by induction that the sequence k_m is decreasing as long as it is defined. Consequently, if the conclusion is false, then

$$\lim_{m\to\infty} k_m = k > 0,$$

and k must satisfy

$$k = 1-\lambda\left(\frac{p+1}{2k}\right)^p.$$

In other words, k is a positive solution to

$$f(x) = x^{p+1}-x^p = -\lambda\left(\frac{p+1}{2}\right)^p.$$

However, the minimum value of $f(x)$ for $x>0$ is easily computed to be

$$-\frac{1}{p+1}\cdot\left(\frac{p}{p+1}\right)^p.$$

Hence we must have

$$\lambda \leqq \left(\frac{2}{p+1}\right)^p \cdot \frac{1}{p+1}\left(\frac{p}{p+1}\right)^p = \lambda_p.$$

Therefore, if $\lambda > \lambda_p$, the conclusion holds.

PROPOSITION 4.11. *Suppose* $n=1$ *and* $q=2p/(p+1)$. *If* $\lambda > \lambda_p$, *then* $z(a)<\infty$ *for all solutions of* (4.2).

Proof. If $z(a)=\infty$, then $u(r)>0$ for all $r>0$. By Proposition 4.4 and the fact that $n=1$, $u(r)$ satisfies conditions (i)-(iii) of Lemma 4.8 with $r_0=0$. Furthermore, by (4.2) with $n=1$, we have that (4.14) holds with $k=1$. Hence by Lemma 4.8, (4.14) and therefore (4.15) hold with all values $k=k_m$ defined in Lemma 4.10. Thus, by Lemma 4.10, $u''(r)\leqq 0$ for all $r>0$. This is impossible since $u'(r)<0$ and $u(r)>0$ for $r>0$.

Proof of Theorem 1.3. Suppose first that $1<q<2p/(p+1)$ and (if $n\geqq 3$) $p<(n+2)/(n-2)$. By Propositions 4.5 and 4.6, for all $\lambda>0$ and $R>0$, there exists $a>0$ such that $z(a)=R$. In other words, if $u(r)$ is the solution to (4.2) with this initial value a, then $u(r)>0$ for $0\leqq r<R$ and $u(R)=0$. Then $\phi(x)=u(|x|)$, $|x|\leqq R$, is the desired solution of (1.5). The other properties of ϕ follow from Corollary 4.3.

Now suppose that $n=1$, $q=2p/(p+1)$, and $\lambda>\lambda_p$. Then by Proposition 4.11 and Lemma 4.7, there exists (a unique) $a>0$ such that $z(a)=R$. The rest of the proof is as in the previous case.

5. Further results on the elliptic problem. In this section we continue in the same context and with the same notation established in the previous section. In particular, for a fixed $\lambda>0$, $z(a)$ is the first zero of the solution to the initial value problem (4.2), with the convention that $z(a)=\infty$ in case the solution always remains positive. Our goal is to study more completely the problem (4.2), i.e., the elliptic problem (1.5) in B_R. We first gather as much information as we can in the general case, and then specialize to dimension $n=1$. The following result is a variation on the estimate (4.6).

LEMMA 5.1. *Let $\lambda>0$ and $u(r)$ be the solution of* (4.2), *with $z(a)$ the first zero of $u(r)$. Then*

$$z(a)\geqq\lambda^{-1/q}a^{1-(p/q)}. \tag{5.1}$$

Proof. We may certainly assume $z(a)<\infty$. We claim first that the maximum value of $-u'(r)$ on $[0, z(a)]$ is achieved in the interior. Indeed,

$$-u''(0)=\frac{\lambda a^p}{n}>0,$$

$$-u''(z(a))=\left(\frac{n-1}{z(a)}\right)u'(z(a))-|u'(z(a))|^q<0.$$

So if r_0 is such that $-u'(r_0)$ is a maximum on $[0, z(a)]$, then $u''(r_0)=0$, i.e.,

$$-(-u'(r_0))^q+\lambda u(r_0)^p=-\frac{n-1}{r_0}u'(r_0)\geqq 0$$

or

$$(-u'(r_0))^q\leqq\lambda u(r_0)^p\leqq\lambda a^p.$$

This implies that for $0\leqq r\leqq z(a)$,

$$-u'(r)\leqq\lambda^{1/q}a^{p/q}.$$

Hence,

$$\begin{aligned}a&=u(0)-u(z(a))\\&=-\int_0^{z(a)}u'(s)\,ds\\&\leqq z(a)\lambda^{1/q}a^{p/q},\end{aligned}$$

which proves (5.1).

LEMMA 5.2. *Let $q>2p/(p+1)$ and (in case $n\geqq 3$) $p<(n+2)/(n-2)$. Then, for all $\lambda>0$, we have*

$$\limsup_{a\to 0}a^{(p-1)/2}z(a)<\infty.$$

In particular, for all sufficiently small a, $z(a)<\infty$ and

$$z(a)\leqq Ca^{-(p-1)/2}. \tag{5.2}$$

Proof. This is an obvious modification to the proof of Proposition 4.6.

Let us consider for a moment whether or not a regular (i.e., C^2) solution of the elliptic problem (1.5) exists on $\Omega = B_R$. As noted in § 4, the methods of [13] can be used to prove that such a solution $\phi(x)$ must be radially symmetric. In other words, $\phi(x) = u(|x|)$, where $u(r)$ is a solution of (4.2) with $z(a) = R$. Therefore, for a fixed $\lambda > 0$, the number of solutions to (1.5) on B_R is precisely the cardinality of the set $z^{-1}(R)$. For each $\lambda > 0$, we define

$$R(\lambda) = \inf_{a>0} z(a). \tag{5.3}$$

Note that if $R(\lambda) < \infty$, then by Proposition 4.5, $z(a) = R$ has at least one solution whenever $R(\lambda) < R < \infty$.

PROPOSITION 5.3. (i) *If $q \geqq p$ then $R(\lambda) > 0$.*

(ii) *If $q > 2p/(p+1)$ and (in case $n \geqq 3$) $p < (n+2)/(n-2)$, then $R(\lambda) < \infty$.*

(iii) *If $q > p$ and (in case $n \geqq 3$) $p < (n+2)/(n-2)$, then $z(a) = R(\lambda) < \infty$ for some $a > 0$, and $z(a) = R$ has at least <u>two</u> solutions $a > 0$ for each $R > R(\lambda)$.*

Proof. Statement (i) follows from the lower estimates for $z(a)$ given by (4.6) and (5.1). Statement (ii) follows from Lemma 5.2. For statement (iii) note that (4.6) and (5.1) imply $\lim_{a\to 0} z(a) = \infty$ and $\lim_{a\to\infty} z(a) = \infty$. Moreover, by Lemma 5.2, $z(a)$ is finite for some values of a, and hence, by Proposition 4.5, assumes its (positive) minimum $R(\lambda)$ at some a_m, $0 < a_m < \infty$. Clearly, then for each $R > R(\lambda)$, there exist a_1 and a_2 with $0 < a_1 < a_m < a_2 < \infty$ such that $z(a_1) = z(a_2) = R$.

COROLLARY 5.4. (i) *If $q \geqq p$ and $0 < R < R(\lambda)$, then there is no regular solution of* (1.5) *on B_R.*

(ii) *If $q > 2p/(p+1)$ and (in case $n \geqq 3$) $p < (n+2)/(n-2)$, then for $R > R(\lambda)$, there is at least one regular solution of* (1.5) *on B_R.*

(iii) *If $q > p$ and (in case $n \geqq 3$) $p < (n+2)/(n-2)$, then for $R = R(\lambda)$, there is at least one regular solution of* (1.5) *on B_R; and for $R > R(\lambda)$, there are at least <u>two</u> regular solutions of* (1.5) *on B_R.*

We next focus our attention on the case $q = 2p/(p+1)$. This is particularly interesting since it is the critical value for both the energy arguments in § 3 and the scaling argument in (the proof of) Proposition 4.6.

PROPOSITION 5.5. *Let $q = 2p/(p+1)$ and $\lambda > 0$. Suppose first $n = 1, 2$ or $n \geqq 3$ and $p < n/(n-2)$. Then there exists a positive constant k such that*

$$U(r) = kr^{-2/(p-1)} \tag{5.4}$$

satisfies

$$u''(r) + \frac{n-1}{r} u'(r) - |u'(r)|^q + \lambda u(r)^p = 0 \tag{5.5}$$

if and only if $\lambda \leqq \lambda_{p,n}$, where

$$\lambda_{p,n} = \frac{1}{p+1}\left[\frac{2p}{(p+1)(2p-np+n)}\right]^p.$$

On the other hand, if $n \geqq 3$ and $p \geqq n/(n-2)$, then such a solution exists for all $\lambda > 0$.

Proof. By direct calculation, we see that $U(r)$, given by (5.4), satisfies (5.5) precisely when

$$\lambda k^{p-1} - \left(\frac{2}{p-1}\right)^q k^{q-1} = -\frac{2}{p-1}\left(\frac{2p}{p-1} - n\right). \tag{5.6}$$

If $n \geqq 3$ and $p \geqq n/(n-2)$, then the right-hand side of (5.6) is nonnegative. In this case, since $q = 2p/(p+1) < p$, a positive solution k to (5.6) can always be found.

Suppose instead either $n=1,2$ or in the case where $n\geqq 3$, $p<n/(n-2)$, so the right-hand side of (5.6) is negative. Then a positive solution k of (5.6) can be found precisely when

$$\inf_{x>0} f(x) \leqq -\frac{2}{p-1}\left(\frac{2p}{p-1}-n\right), \tag{5.7}$$

where $f(x)=\lambda x^{p-1}-[2/(p-1)]^q x^{q-1}$. Using elementary calculus, and not forgetting that $q=2p/(p+1)$, we can verify that (5.7) holds if and only if $\lambda \leqq \lambda_{p,n}$.

Remark. Note that $\lambda_{p,1}=\lambda_p$, defined by (1.6). Also, $\lambda_{p,n}$ is increasing as a function of n.

PROPOSITION 5.6. *Assume* $q=2p/(p+1)$ *and* $\lambda \leqq \lambda_p=\lambda_{p,1}$. *Then* $z(a)=\infty$ *for all* $a>0$. *In other words, there is no regular solution of* (1.5) *on* B_R, *for any* $R>0$.

Proof. In dimension $n=1$ there is a particularly easy and elegant proof, which we present first. Suppose a C^2 solution ϕ of (1.5) exists in dimension $n=1$. Let $U(r)$ be the solution of (5.5) given by (5.4) with $n=1$. (Recall $\lambda \leqq \lambda_{p,1}$.) Set

$$\bar{\rho}=\sup\{\rho\in\mathbf{R}\text{: the graph of } \phi(r-\rho) \text{ does not touch the graph of } U(r)\}.$$

Clearly, $\bar{\rho}\in\mathbf{R}$. Also, the graphs of $\phi(r-\bar{\rho})$ and $U(r)$ touch at some point r_0, i.e., $\phi(r_0-\bar{\rho})=U(r_0)>0$; and by the definition of $\bar{\rho}$, we must also have $\phi'(r_0-\bar{\rho})=U'(r_0)$. However, both $\phi(r-\bar{\rho})$ and $U(r)$ satisfy (5.5) with the same Cauchy data at r_0. Hence $\phi(r-\bar{\rho})=U(r)$ wherever both functions are defined. In particular, ϕ cannot be $C^2([-R,R])$ with $\phi(\pm R)=0$.

For the case $n\geqq 2$, we assume that $z(a)<\infty$ for some $a>0$. For any fixed $\gamma>0$, let

$$G(r)=\frac{u'(r)^2}{2}-\gamma|u(r)|^p u(r).$$

Then $G(0)<0$ and $G(z(a))>0$. ($u'(z(a))=0$ by local existence and uniqueness.) Thus, the first zero of $G(r)$ is between zero and $z(a)$; call it r_0. Clearly, $G(r_0)=0$ and $G'(r_0)\geqq 0$. However, since $G(r_0)=0$, we have

$$|u'(r_0)|=\sqrt{2\gamma}\,u(r_0)^{(p+1)/2}.$$

Therefore, at $r=r_0$

$$\begin{aligned} G' &= u'u''-\gamma(p+1)u^p u' \\ &= u'\left(-\frac{n-1}{r_0}u'+|u'|^q-\lambda u^p-\gamma(p+1)u^p\right) \\ &< u'(|u'|^q-(\lambda+\gamma(p+1))u^p) \\ &= u'((2\gamma)^{q/2}u^{(p+1)q/2}-(\lambda+\gamma(p+1))u^p) \\ &= u'u^p f(\gamma), \end{aligned}$$

where

$$f(\gamma)=(2\gamma)^{p/(p+1)}-\gamma(p+1)-\lambda.$$

(We have used the fact that $(p+1)q/2=p$.) Now $G'(r_0)\geqq 0$, $u'(r_0)<0$, and $u(r_0)>0$. Consequently, we must have $f(\gamma)<0$. Since $\gamma>0$ was arbitrary, this must be true for all $\gamma>0$. A straightforward calculation of the extreme points for $f(\gamma)$ shows that we must have $\lambda>\lambda_{p,1}$. This proves the proposition.

Remarks. The estimate for $G'(r_0)$ depends on the fact that $n \geqq 2$ in order to get strict inequality. Thus, it seems that for $n=1$, the second argument misses the case $\lambda = \lambda_{p,1}$. However, since for a fixed value of a, the set of $\lambda > 0$ for which $z(a) < \infty$ is clearly open, we recover the case $\lambda = \lambda_{p,1}$ when $n = 1$.

Also, in the case $n=1$, Proposition 5.6 and Theorem 1.3 give a complete description of when there are solutions of (1.5) on B_R with $q = 2p/(p+1)$. There exists a solution if and only if $\lambda > \lambda_p$. It is natural to conjecture an analogous result for higher dimensions. We wonder what the sharp cutoff value would be, in particular if it is $\lambda_{p,n}$.

A variation on the proof of Proposition 5.6 yields the following result.

PROPOSITION 5.7. *Assume $q < 2p/(p+1)$ and fix $\lambda > 0$. Then there exists $a_* > 0$ such that $z(a) = \infty$ for $a \leqq a_*$.*

Proof. Suppose $z(a) < \infty$. Let $u(r)$ be the corresponding solution of (4.2) and set

$$G(r) = \frac{u'(r)^2}{2} - \lambda |u(r)|^p u(r).$$

Then $G(0) < 0$ and $G(z(a)) > 0$. Let r_0 be the first zero of G; so $G(r_0) = 0$ and $G'(r_0) \geqq 0$. Reasoning as in the proof of Proposition 5.6 (with $\gamma = \lambda$), we see that at $r = r_0$

$$\begin{aligned} G' &\leqq u'((2\lambda)^{q/2} u^{(p+1)q/2} - \lambda(p+2)u^p) \\ &\leqq u' u^{(p+1)q/2}((2\lambda)^{q/2} - \lambda(p+2)a^{p-[(p+1)q/2]}). \end{aligned}$$

Since $p > (p+1)q/2$, it follows that for a sufficiently small, $G'(r_0) < 0$. This contradicts the earlier observation that $G'(r_0) \geqq 0$.

Hence $z(a) = \infty$ for $a > 0$ sufficiently small.

We now restrict ourselves to the special case $n = 1$. The problem (4.2) becomes the autonomous problem

$$\begin{aligned} &u''(r) - |u'(r)|^q + \lambda |u(r)|^p = 0, \\ &u(0) = a > 0, \qquad u'(0) = 0. \end{aligned} \tag{5.8}$$

Problem (1.5) becomes

$$\begin{aligned} &\phi'' - |\phi'|^q + \lambda \phi^p = 0 \quad \text{in } (-R, R), \\ &\phi > 0 \quad \text{in } (-R, R), \\ &\phi(\pm R) = 0, \end{aligned} \tag{5.9}$$

where $\phi \in C^2([-R, R])$.

Consider the case $q < 2p/(p+1)$. By Propositions 4.5 and 4.6, i.e., the first part of Theorem 1.3, for all $\lambda > 0$ and $R > 0$, there is a solution to problem (5.9). We will show it to be unique. (Note that in the case $q = 2p/(p+1)$ the solution of (5.9) is unique when it exists because of formula (4.13). If $q > p$, we know it is not unique for R large enough.)

LEMMA 5.8. *Let $v(r)$ be the maximal solution (as in Proposition 4.4) to the problem*

$$\begin{aligned} &v''(r) - b|v'(r)|^q + \lambda |v(r)|^p = 0, \\ &v(0) = v_0 > 0, \qquad v'(0) = 0, \end{aligned} \tag{5.10}$$

where $b > 0$ and $\lambda > 0$ are parameters. Then $v(r)$ is an increasing function of b.

Proof. The existence and uniqueness of $v(r)$ follow exactly as in the proof of Proposition 4.1. In particular, $v'(r) < 0$ for $r < 0$. It is clear from the integral equation corresponding to (5.10) that v is a C^1 function jointly in r and b. We denote $v_r = \partial v/\partial r$

and $v_b = \partial v/\partial b$. Then considering v_r and v_b as functions of r, and using $'$ to denote d/dr, we have

$$v_r'' + bq|v'|^{q-1}v_r' + \lambda p|v|^{p-1}v_r = 0,$$
$$v_b'' + bq|v'|^{q-1}v_b' + \lambda p|v|^{p-1}v_r = (-v')^q.$$

Hence, setting $w = v_b'v_r - v_r'v_b$, it follows that

$$w' + qb(-v')^{q-1}w = (-v')^q v' < 0,$$

or, if f denotes a primitive of $qb(-v')^{q-1}$, that

$$(e^f w)' < 0. \tag{5.11}$$

Since $v_b(0) = v_r(0) = 0$, we have $w(0) = 0$. Hence by (5.11), $w(r) < 0$ for $r > 0$. This implies $(v_b/v_r)' < 0$ for $r > 0$. Moreover,

$$\lim_{r\to 0} \frac{v_b(r)}{v_r(r)} = \lim_{r\to 0} \frac{v_b'(r)}{v_r'(r)} = \frac{0}{v''(0)} = 0,$$

and so $v_b/v_r < 0$ for all $r > 0$, i.e., $v_b(r) > 0$ for $r > 0$.

PROPOSITION 5.9. *Let $q < 2p/(p+1)$ and $n = 1$. Then, for all $\lambda > 0$, $a^{(p-1)/2}z(a)$ is a decreasing function of a. In particular, $z(a)$ is decreasing.*

If $q > 2p/(p+1)$ and $n = 1$, then for all $\lambda > 0$, $a^{(p-1)/2}z(a)$ is an increasing function of a.

Proof. By the previous lemma, if $n = 1$ and $q < 2p/(p+1)$, then $v_a(r)$ defined by (4.9) is a decreasing function of a. Consequently, the first zero of v_a, i.e., $a^{(p-1)/2}z(a)$, is a decreasing function of a. More precisely, $a^{(p-1)/2}z(a)$ is nonincreasing for all $a > 0$ and strictly decreasing where it is finite.

An analogous argument works for $q > 2p/(p+1)$.

COROLLARY 5.10. *Let $q < 2p/(p+1)$ and $n = 1$. Then for every $\lambda > 0$ and $R > 0$, the C^2 solution of (5.9) is unique.*

Finally, we have a result that further contrasts the cases $q < 2p/(p+1)$ and $q > 2p/(p+1)$.

PROPOSITION 5.11. *Suppose $q > 2p/(p+1)$ and $n = 1$. Let $\lambda > 0$ be arbitrary. Then $z(a) < \infty$ for all $a > 0$ and $\lim_{a\to 0} z(a) = \infty$.*

Proof. By Lemma 5.2 and Proposition 4.5 we already know that $z(a) < \infty$ for small $a > 0$ and $\lim_{a\to 0} z(a) = \infty$. Suppose $z(a) = \infty$ for some $a > 0$, and let $u(r)$ be the corresponding solution of (5.8). Let $v(r)$ be a solution to (5.8) with a smaller initial value a such that $z(a) < \infty$. Let

$$\bar{\rho} = \sup\{\rho \in \mathbf{R}: \text{the graph of } v(r-\rho) \text{ does not touch the graph of } u(r)\}.$$

As in the proof of Proposition 5.6, it is clear that $\bar{\rho} \in R$ and that $u(r)$ and $v(r-\bar{\rho})$ would have to coincide, which is impossible.

Remarks. We can make a few more observations in the case $n = 1$. First, if $q < 2p/(p+1)$, then by Propositions 4.6, 5.7, and 5.9 there exists $a_* > 0$ such that $z(a) = \infty$ for $a \leqq a_*$ and $z(a) < \infty$ for $a > a_*$. Next, in the case $q > 2p/(p+1)$, if ϕ_1 and ϕ_2 are two different solutions of (5.9), then $\phi_1(x) \neq \phi_2(x)$ for all x in $(-R, R)$. This follows from a translation argument similar to the proofs of Propositions 5.6 and 5.11. We mention without proof that if $q > 2p/(p+1)$, there can exist singular solutions of (5.9), i.e., solutions in $C^2([-R, R]\backslash\{0\})$ with $\lim_{x\to 0} \phi(x) = \infty$.

Clearly, solutions of (1.5) exhibit radically different behavior depending on the relationship between p and q. However, the picture is certainly not complete.

Note added in proof. After we completed this paper, we learned that B. Kawohl and L. Peletier obtained, among other interesting results, blowup in the case $q = 2$. (See B. Kawohl and L. Peletier [36].)

Acknowledgments. While at the Institute for Mathematics and Its Applications (IMA), the University of Minnesota, the authors were fortunate to be able to discuss this work with S. Hastings, R. Pego, and P. Souganidis. We thank them for their ideas and suggestions, and we are grateful to the IMA for its support.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] J. M. BALL, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math., Oxford Ser. (2), 28 (1977), pp. 473-486.
[3] P. BARAS AND L. COHEN, *Complete blow-up after $T_{\max}$ for the solution of a semilinear heat equation*, J. Funct. Anal., 71 (1987), pp. 142-174.
[4] J. BEBERNES, A. BRESSAN, AND D. EBERLY, *A description of blow-up for the solid fuel ignition model*, Indiana Univ. Math. J., 36 (1987), pp. 295-305.
[5] J. BEBERNES AND W. FULKS, *The small heat-loss problem*, J. Differential Equations, 57 (1985), pp. 324-332.
[6] J. BEBERNES AND D. R. KASSOY, *A mathematical analysis of blow-up for thermal reaction—the spatially nonhomogeneous case*, SIAM J. Appl. Math., 40 (1981), pp. 476-484.
[7] J. BEBERNES AND W. TROY, *On the nonexistence for the Kassoy problem in dimension* 1, SIAM J. Math. Anal., 18 (1987), pp. 1157-1162.
[8] H. BELLOUT, *A criterion for blow-up solution to semi-linear heat equations*, SIAM J. Math. Anal., 18 (1987), pp. 722-727.
[9] H. BREZIS, *Analyse fonctionnelle*, Masson, Paris, 1983.
[10] H. ENGLER, *Contractive properties for the heat equation in Sobolev spaces*, preprint.
[11] A. FRIEDMAN AND A. A. LACEY, *Blow-up time for solutions of nonlinear heat equations with small diffusion*, SIAM J. Math. Anal., 18 (1987), pp. 711-721.
[12] A. FRIEDMAN AND J. B. MCLEOD, *Blow-up of solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425-448.
[13] B. GIDAS, W.-M. NI, AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209-243.
[14] Y. GIGA, *A bound for global solutions of semilinear heat equations*, Comm. Math. Physics, 103 (1986), pp. 415-421.
[15] Y. GIGA AND R. V. KOHN, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297-319.
[16] ———, *Characterizing blow-up using similarity variables*, Indiana Univ. Math. J., 36 (1987), pp. 1-38.
[17] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.
[18] A. HARAUX AND F. B. WEISSLER, *Non-uniqueness for a semilinear initial value problem*, Indiana Univ. Math. J., 31 (1982), pp. 167-189.
[19] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[20] T. KATO AND H. FUJITA, *On the nonstationary Navier-Stokes system*, Rend. Sem. Mat. Univ. Padova, 32 (1962), pp. 243-260.
[21] H. KOMATSU, *Fractional powers of operations*, Pacific J. Math., 19 (1966), pp. 285-346.
[22] A. A. LACEY, *Mathematical analysis of thermal runaway for spatially inhomogeneous reactions*, SIAM J. Appl. Math., 43 (1983), pp. 1350-1366.
[23] ———, *The form of blow-up for nonlinear parabolic equations*, Proc. Roy. Soc. Edinburgh Sect. A, 98 (1984), pp. 183-202.
[24] ———, *Global blow-up of a nonlinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 104 (1986), pp. 161-167.
[25] C. E. MUELLER AND F. B. WEISSLER, *Single point blow-up for a general semilinear heat equation*, Indiana Univ. Math. J., 34 (1985), pp. 881-913.
[26] R. SEELEY, *Norms and domains of complex powers A_B^z*, Amer. J. Math., 93 (1971), pp. 299-309.
[27] ———, *Interpolation in L^p with boundary conditions*, Stud. Math., 44 (1972), pp. 47-60.
[28] I. SEGAL, *Nonlinear semi-groups*, Ann. of Math., 78 (1963), pp. 339-364.

[29] W. C. TROY, *The existence of bounded solutions of a semilinear heat equation*, preprint.
[30] W. C. TROY AND D. EBERLY, *On the existence of a logarithmic type solution to the Kassoy-Kapila problem in dimension* $3 \leqq n \leqq q$, J. Differential Equations, to appear.
[31] F. B. WEISSLER, *Semilinear evolution equations in Banach spaces*, J. Funct. Anal., 32 (1979), pp. 277-296.
[32] ———, *Single point blow-up for a semilinear initial value problem*, J. Differential Equation, 55 (1984), pp. 204-224.
[33] ———, *An* L^∞ *blow-up estimate for a nonlinear heat equation*, Comm. Pure Appl. Math., 38 (1985), pp. 291-295.
[34] ———, *Local existence and nonexistence for semilinear parabolic equations in* L^p, Indiana Univ. Math. J., 29 (1980), pp. 79-102.
[35] H. A. LEVINE, *Some nonexistence and instability theorems for solutions of formally parabolic equations of the form* $Pu_t = -Au + F(u)$, Arch. Rational Mech. Anal., 51 (1973), pp. 371-386.
[36] B. KAWOHL AND L. PELETIER, *Observation on blow-up and dead cores for nonlinear parabolic equations*, to appear.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 908-929, July 1989

008

THE RIEMANN PROBLEM FOR MULTICOMPONENT POLYMER FLOODING*

THORMOD JOHANSEN† AND RAGNAR WINTHER‡

Abstract. The global Riemann problem for a nonstrictly hyperbolic system of conservation laws modeling multicomponent polymer flooding is solved. The solution is constructed by first generating the Riemann solution of an associated one-phase problem.

Key words. Riemann problems, polymer flooding, adsorption

AMS(MOS) subject classifications. 35L65, 76S05

1. Introduction. In this paper we construct the global solution of the Riemann problem for a nonstrictly hyperbolic system of conservation laws of the following form:

$$\begin{aligned} &s_t + f(s, c_1, \cdots, c_n)_x = 0, \\ &[sc_i + a_i(c_i)]_t + [c_i f(s, c_1, \cdots, c_n)]_x = 0, \qquad i = 1, 2, \cdots, n, \end{aligned} \tag{1.1}$$

where the unknown state vector $(s, c_1, \cdots, c_n) = (s, c) \in \mathbb{I}^{n+1}$ and $f: \mathbb{I}^{n+1} \to \mathbb{R}$ and $a_i: \mathbb{I} \to \mathbb{R}$ are given smooth functions. Here $\mathbb{I}$ denotes the unit interval $\mathbb{I} = [0, 1]$.

The results presented here generalize the results of Johansen and Winther [9] where the general Riemann problem for the model (1.1) is solved for $n = 1$.

The main purpose of this paper is to extend this solution to the case where $n > 1$. We show that a certain "projection principle" makes it possible to construct the Riemann solution in this case from a coupled sequence of Riemann problems of the form studied in [9]. This construction is the main result of the paper.

The system (1.1) models simultaneous one-dimensional flow of two immiscible phases (the aqueous phase and the oilic phase) in a homogeneous porous medium. It is further assumed that n chemical components are dissolved in the aqueous phase. These components could, for example, be different polymers that all have different influence on the flow properties. The equations are derived from conservation of mass of the two phases and of the n chemical components. The variable s denotes the saturation of the aqueous phase. The term a_i in (1.1) models the adsorption of component i on the porous medium. It is assumed throughout the paper that a_i is a given function of concentration that depends only on the variable c_i; i.e., $a_i = a_i(c_i)$. As in [9] we will assume that each of the adsorption functions a_i are of Langmuir type; i.e., a_i is a concave, increasing function such that $a_i(0) = 0$. No relation between the different adsorption functions will be assumed. The function $f = f(s, c_1, \cdots, c_n)$ is the fractional flow function of the aqueous phase. We will assume that this function is a decreasing function with respect to each of the variables c_i. This corresponds, for example, to the effect that the viscosity of the aqueous phase is increasing with respect to the concentration c_i. For any fixed concentration vector c the function f is assumed to be an increasing function of s with one inflection point. For more details on the physical derivation of the model (1.1) we refer to [9] and references given therein.

* Received by the editors November 4, 1987; accepted for publication (in revised form) October 4, 1988. This work was supported by Vista, a research cooperation between the Norwegian Academy of Science and Letters and Statoil.

† Institute for Energy Technology, P.O. Box 40, 2007 Kjeller, Norway. The work of this author was supported in part by the Royal Council for Scientific and Industrial Research, Norway.

‡ Institute of Informatics, University of Oslo, P.O. Box 1080 Blindern, 0316 Oslo 3, Norway.

A model, where the effects of polymer and surfactants are modeled as described above, is for example proposed by Lake and Helfferich [11]. A more realistic description of the adsorption effects would be that the different components compete for the vacant adsorption sites. In [13], Rhee, Aris, and Amundson solve the Riemann problem for such a model in a one-phase situation. They assume that the adsorption effects are of generalized Langmuir type; i.e., the adsorption functions a_i are of the form

$$a_i = a_i(c_1, c_2, \cdots, c_n) = K_i c_i \Big/ \left(1 + \sum_{j=1}^{n} K_j c_j\right) \tag{1.2}$$

for $i = 1, 2, \cdots, n$, where $K_1 < K_2 < \cdots < K_n$ are given positive constants. The techniques developed in this paper can also be used to construct the Riemann solution for a two-phase model of the form (1.1), but where the functions $a_i(c_i)$ are replaced by adsorption functions of the form (1.2). This construction will be discussed in a forthcoming paper.

The model (1.1) is an example of a system of hyperbolic conservation laws. However, under the present assumptions on the fractional flow function f, the system is not strictly hyperbolic; i.e., there exist regions in the state space where the eigenvalues of the appropriate Jacobian matrix are not distinct.

The Riemann problem for the model (1.1) consists of constructing a weak solution of the pure initial value problem for (1.1) with initial condition

$$(s, c)(x, 0) = \begin{cases} (s^L, c^L) & \text{if } x < 0, \\ (s^R, c^R) & \text{if } x > 0, \end{cases} \tag{1.3}$$

where the left and right states (s^L, c^L) and (s^R, c^R) in I^{n+1} are arbitrary. To distinguish the physically meaningful weak solutions we will also require that any discontinuity of the solution satisfies an "entropy condition" obtained from traveling wave analysis. As in [9] the entropy condition will allow overcompressive shocks. In particular, this condition will guarantee the uniqueness of the solution of the Riemann problem.

The interest in Riemann problems is partly motivated from their potential application as building blocks in the construction of numerical methods. Examples of such methods are the Random Choice Method [1]-[3], Godunov-type methods [6] and front-tracking techniques [4], [5]. It is therefore of interest to develop a computer program that computes the desired solution of the Riemann problem. Such a program has been implemented from the constructive existence proof given in this paper. In a forthcoming paper we will discuss this implementation and also present some examples of solutions.

If $n = 1$ and if the adsorption term $a_i(c_i)$ is neglected, the model (1.1) corresponds to a 2×2 system of conservation laws analyzed by Keyfitz and Kranzer [10] and Isaacson [7]. In [9] the adsorption term is included. The effect of this term is that the linearly degenerate characteristic field appearing in the analysis of [7] and [10] is replaced by a nondegenerate field. A further consequence of the adsorption term is that the state-space solution of the Riemann problem becomes unique. A multicomponent version of the nonadsorption model of [7] and [10] is studied by Isaacson and Temple [8]. In the present paper the model (1.1), with adsorption terms included, is analyzed in the case when $n > 1$. The results depend heavily on the results of [9] and on the fact that the Riemann problem decouples according to a "projection principle." As a consequence of this principle the Riemann problem for (1.1) can be constructed by first solving the Riemann problem for an associated one-phase problem. Thereafter, the solution for (1.1) can be obtained from a finite sequence of coupled 2×2 Riemann problems.

The precise assumptions on the system (1.1) are stated in § 2. In § 3 the properties of the elementary waves (rarefactions and shocks) are described. The "projection principle" is discussed in § 4, while the complete construction of the Riemann solution is performed in § 5. The main conclusions obtained from this paper are given in § 6.

2. A precise formulation of the model. The model (1.1) is a system of $n+1$ hyperbolic conservation laws with $n+1$ unknowns $s, c_1, \cdots, c_n$. We frequently let c denote the n-vector $(c_1, c_2, \cdots, c_n)$. For the partial derivatives of the flux function $f=f(s, c)=f(s, c_1, \cdots, c_n)$ we use the notation

$$f_s=\frac{\partial f}{\partial s} \quad \text{and} \quad f_i=\frac{\partial f}{\partial c_i} \quad \text{for } i=1, 2, \cdots, n.$$

Furthermore, f_c denotes the row vector $f_c=(f_1, f_2, \cdots, f_n)$.

The derivatives of the adsorption functions $a_i=a_i(c_i)$ are denoted by $h_i=da_i/dc_i$, and $H(c)$ is the $n\times n$ matrix

$$H(c)=\operatorname{diag}(h_1(c_1), \cdots, h_n(c_n)).$$

The assumptions on the functions f and a_i are similar to those made in [9]. In particular, the adsorption functions a_i are assumed to be smooth, strictly increasing, strictly concave functions of c_i such that $a_i(0)=0$ (cf. Fig. 2.1); i.e.,

$$h_i(c_i)>0 \quad \text{and} \quad \frac{dh_i}{dc_i}(c_i)<0 \quad \text{for } c_i\in\mathbb{I}. \tag{2.1}$$

The flux function f is assumed to be a smooth function such that (cf. Fig. 2.2):

(2.2a) $f(0, c)\equiv 0, \quad f(1, c)\equiv 1,$

(2.2b) $f_s(s, c)>0$ for $0<s<1$, $c\in\mathbb{I}^n$,

(2.2c) $f_i(s, c)<0$ for $0<s<1$, $c\in\mathbb{I}^n$,

(2.2d) for each $c\in\mathbb{I}^n$, $f(\cdot, c)$ has a unique point of inflection $s^I=s^I(c)\in\mathbb{I}$ such that $f_{ss}(s, c)>0$ for $s\in(0, s^I)$, $f_{ss}(s, c)<0$ for $s\in(s^I, 1)$.

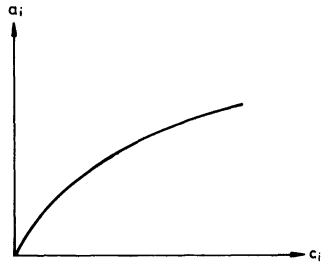


FIG. 2.1

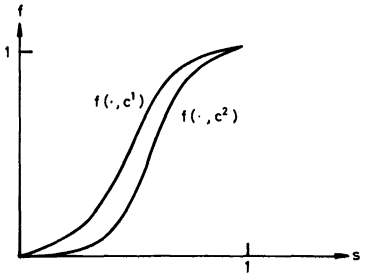


FIG. 2.2. $c_j^1=c_j^2$ *for* $j\neq i$, $c_i^1<c_i^2$.

Let $u \in \mathbb{R}^{n+1}$ denote the state vector $u = (s, c)$. If the solution of the system (1.1) is smooth, then the system can be rewritten in the form

$$u_t + A(u)u_x = 0,$$

where $A(u)$ is the $(n+1)\times(n+1)$ matrix

$$A(u) = \begin{pmatrix} f_s(s, c) & f_c(s, c) \\ 0 & f(s, c)(sI + H(c))^{-1} \end{pmatrix}.$$

Here I denotes the $n \times n$ identity matrix. The eigenvalues of A are $\lambda_s = f_s$ and $\lambda_i = f/(s + h_i)$. In general these eigenvalues are not distinct.

For a given $c \in \mathbb{I}^n$ there is at most one value of $s > 0$ such that $\lambda_s = \lambda_i$ (cf. Fig. 2.3). In particular, for constant values of c_j, for $j \neq i$, the set $\{(s, c) \mid \lambda_s(s, c) = \lambda_i(s, c), s > 0\}$ defines a curve in (s, c_i)-space similar to the transition curve T appearing in the analysis of [9]. We impose the same regularity requirements on all of these curves as we did on the corresponding curve in [9]; i.e., for arbitrary constant values of c_j, for $j \neq i$, there is a unique value c_i^T such that the curve $\lambda_i = \lambda_s$ exists for $0 \leqq c_i \leqq c_i^T$ (cf. Fig. 2.4).

Furthermore, the relation $\lambda_i = \lambda_j$ defines an n-dimensional surface in state space given by

$$h_j(c_j) = h_i(c_i) \quad \text{or} \quad c_j = h_j^{-1}(h_i(c_i)) \equiv g_{i,j}(c_i).$$

We observe that it follows from the assumptions on the functions h_i that the functions $g_{i,j}$ are increasing functions of c_i (cf. Fig. 2.5).

3. Elementary waves. In this section we determine all the elementary waves of the model (1.1); i.e., we determine the rarefaction waves and the shock waves. The

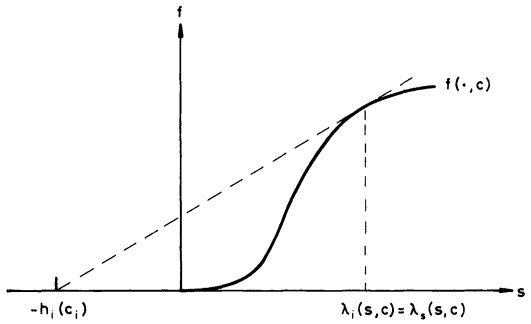


FIG. 2.3. $\lambda_i = \lambda_s$.

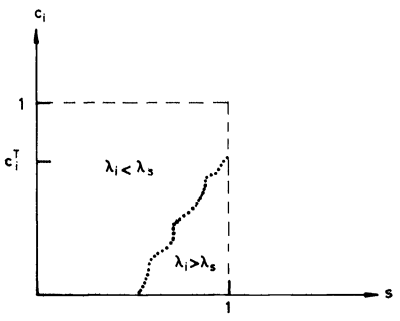


FIG. 2.4

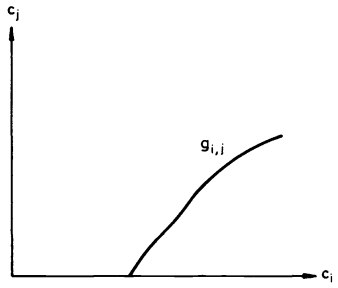


FIG. 2.5

construction of the solution of the general Riemann problem will thereafter be performed by composing different elementary waves and constant states.

Any weak solution of the pure initial value problem for (1.1), with initial conditions of the form (1.3), depends only on the variable x/t. The rarefaction waves are the possible smooth solutions. Let $u(x, t) = (s, c)(x, t) = v(x/t)$ be a smooth solution of (1.1), (1.2). It is well known that the values of v must lie on an integral curve of one of the (right) eigenvectors of the matrix $A(v)$, while $\xi = x/t$ is the corresponding eigenvalue $\lambda(v)$. Thus $\lambda(v(\xi))$ must increase monotonically with ξ. The requirement gives a direction, corresponding to increasing values of $\xi = x/t$, associated with the integral curves. These directed integral curves are called *rarefaction curves.*

To determine the rarefaction curves we have to find the eigenvectors of the matrix $A(v)$. If the eigenvalue $\lambda = \lambda_s = f_s$ then the corresponding eigenvector r is given by $r = (1, 0)$ (i.e., $c_i = 0$ for $i = 1, 2, \cdots, n$). Hence, the associated integral curves are the curves c-constant. The associated rarefaction waves correspond to rarefaction waves of the single Buckley–Leverett equation

$$s_t + f(s)_x = 0, \tag{3.1}$$

where $f(s) = f(s, c)$, $c =$ constant. These waves are referred to as *s-rarefactions.*

If $\lambda = \lambda_i$ for some i then the vector c will be nonconstant along the associated integral curves. These rarefaction waves will therefore be referred to as *c-rarefactions.* An eigenvector corresponding to $\lambda = \lambda_i$ is given by

$$r_i = f_i(1, 0) + (\lambda_i - \lambda_s)(0, e_i), \tag{3.2}$$

where e_i denotes the ith unit vector in $\mathbb{R}^n$ and $(1, 0)$ is the first unit vector in $\mathbb{R}^{n+1}$.

Hence, the associated integral curves are determined by

$$f_i \frac{dc_i}{ds} = \lambda_i - \lambda_s, \qquad c_j = \text{constant for } j \neq i, \tag{3.3}$$

where $c_i = c_i(s)$. Since $f_i(s, c) < 0$ for $0 < s < 1$ this implies that $dc_i/ds > 0$ when $\lambda_s > \lambda_i$ and that $dc_i/ds < 0$ when $\lambda_s < \lambda_i$. Furthermore, a straightforward calculation shows that

$$\frac{d\lambda_i}{ds} = -\frac{\lambda_i}{s + h_i} \frac{dh_i}{dc_i} \frac{dc_i}{ds}$$

along the integral curves. Since $dh_i/dc_i < 0$, this implies that the rarefaction curves are directed toward increasing values of c_i (cf. Fig. 3.1). If $\lambda = \lambda_i = \lambda_j$ for $i \neq j$ then any linear combination of the eigenvectors r_i and r_j given by (3.2) corresponds to an eigenvector associated with this eigenvalue. Furthermore, if such an eigenvector is

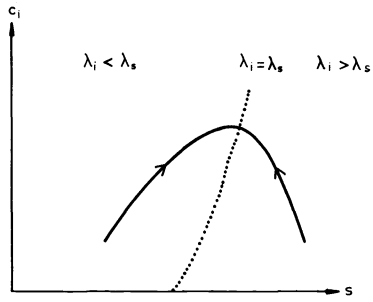


FIG. 3.1

tangential to the surface $\lambda_i = \lambda_j$, the corresponding integral curve is a curve on the surface $\lambda_i = \lambda_j$. Hence, if we choose

$$r = \left(f_i + f_j \frac{dg_{i,j}}{dc_i}\right)(1, 0) + (\lambda - \lambda_s)\left[(0, e_i) + \frac{dg_{i,j}}{dc_i}(0, e_j)\right]$$

we obtain the integral curve

$$\begin{aligned} &\left(f_i + f_j \frac{dg_{i,j}}{dc_i}\right)\frac{dc_i}{ds} = \lambda - \lambda_s, \\ &c_j = g_{i,j}(c_i), \\ &c_k = \text{constant for } k \neq i, j, \end{aligned} \tag{3.4}$$

where $c_i = c_i(s)$. This is a curve on the surface $\lambda_i = \lambda_j$. Since $f_i + f_j(dg_{i,j}/dc_i) < 0$ for $0 < s < 1$, the same analysis as above shows that the projection of the integral curves into (s, c_i)-space have the form illustrated in Fig. 3.1. In particular, the variable c_i is increasing in the direction of the rarefaction curves.

The rarefaction curves generated by (3.4), where $\lambda_i = \lambda_j$, will be referred to as c-rarefactions of multiplicity two. The analysis above can easily be generalized to *c-rarefactions of multiplicity $m+1$* on the surface $\lambda_i = \lambda_{j_1} = \lambda_{j_2} = \cdots = \lambda_{j_m}$. These curves are determined by the system

$$\begin{aligned} &\left(f_i + \sum_{j \in J} f_j \frac{dg_{i,j}}{dc_i}\right)\frac{dc_i}{ds} = \lambda - \lambda_s, \\ &c_j = g_{i,j}(c_i) \quad \text{for } j \in J, \\ &c_k = \text{constant} \quad \text{for } k \neq i \text{ and } k \notin J, \end{aligned}$$

where J denotes the index set $J = \{j_1, j_2, \cdots, j_m\}$. As above these rarefaction curves are directed toward increasing values of c_i (and $c_{j_1}, \cdots, c_{j_m}$), and the projection of the curves into (s, c_i)-space have the form illustrated by Fig. 3.1.

We next determine the shock waves of the model (1.1); i.e., for given states $u^L, u^R \in \mathbb{I}^{n+1}$ we derive possible weak solutions of (1.1) of the form

$$u(x, t) = \begin{cases} u^L & \text{if } x/t < \sigma, \\ u^R & \text{if } x/t > \sigma, \end{cases} \tag{3.5}$$

where σ is the shock speed. We will also require that the shock waves satisfy an "entropy condition."

Any weak solution of (1.1) of the form (3.5) must satisfy the Rankine-Hugoniot condition given by

$$
\begin{aligned}
&f(s^R, c^R) - f(s^L, c^L) = \sigma(s^R - s^L),\\
&c_i^R f(s^R, c^R) - c_i^L f(s^L, c^L) = \sigma[s^R c_i^R + a_i(c_i^R) - s^L c_i^L - a_i(c_i^L)]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } i = 1, 2, \cdots, n.
\end{aligned} \tag{3.6}
$$

If $c^R = c^L$ (i.e., $c_i^R = c_i^L$ for $i = 1, 2, \cdots, n$), then (3.6) reduces to the single equation

$$f(s^R, c) - f(s^L, c) = \sigma(s^R - s^L), \tag{3.7}$$

where $c = c^R = c^L$. This corresponds to the Rankine-Hugoniot condition for the single Buckley-Leverett equation (3.1). Corresponding to the theory for scalar conservation laws, shock waves of this form satisfy an entropy condition if and only if

$$[f(s, c) - f(s^L, c) - \sigma(s - s^L)] \operatorname{sign}(s - s^L) \geqq 0 \tag{3.8}$$

for any s between s^L and s^R (cf. [9], [12], [15] and references given therein). These shocks, with $c = c^R = c^L$ and which satisfy (3.7) and (3.8), are referred to as *s-shocks.*

Consider the Rankine-Hugoniot relations (3.6) when $c^L \neq c^R$. Shock waves of this form will be referred to as *c-shocks.* If $c_i^L \neq c_i^R$ for some i, and $c_j^L = c_j^R$ for $j \neq i$, the relations (3.6) can be written in the following form (cf. [9]):

$$\frac{f(s^R, c^R)}{s^R + h_i^L(c_i^R)} = \frac{f(s^L, c^L)}{s^L + h_i^L(c_i^R)} = \sigma, \tag{3.9}$$

where $h_i^L(c_i)$ is defined by

$$h_i^L(c_i) = \begin{cases} (a_i(c_i) - a_i(c_i^L))/(c_i - c_i^L) & \text{if } c_i \neq c_i^L, \\ h_i(c_i) & \text{if } c_i = c_i^L. \end{cases}$$

We observe that the value $h_i^L(c_i^R)$ is determined from the values of c_i^L and c_i^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 relation (3.9) (cf. Fig. 3.2).

In [9] the entropy conditions for the c-shocks have been derived from a traveling wave analysis. Here, it suffices to simply state the result of a corresponding analysis. For the c-shocks, with $c_i^L \neq c_i^R$ and $c_j^L = c_j^R$ for $j \neq i$, the entropy conditions are

$$c_i^L > c_i^R \tag{3.10}$$

and

$$\lambda_s(u^R) < \sigma \quad \text{or} \quad \lambda_s(u^L), \lambda_s(u^R) \geqq \sigma. \tag{3.11}$$

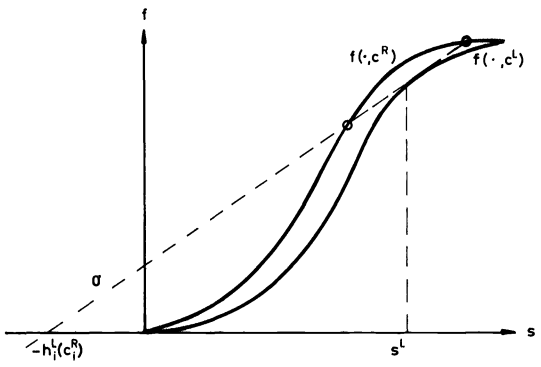


FIG. 3.2

Since the function h_i is strictly decreasing, relation (3.9) implies that (3.10) is equivalent to the eigenvalue/shock speed relation

$$\lambda_i(u^L) > \sigma > \lambda_i(u^R).$$

Hence, if in addition

$$\lambda_s(u^L) \geqq \sigma > \lambda_s(u^R),$$

then we do allow overcompressive shocks where both characteristics on both sides of the shock enter the shock (cf. Schaeffer and Schearer [14]). An example of such a shock wave is illustrated in Fig. 3.3.

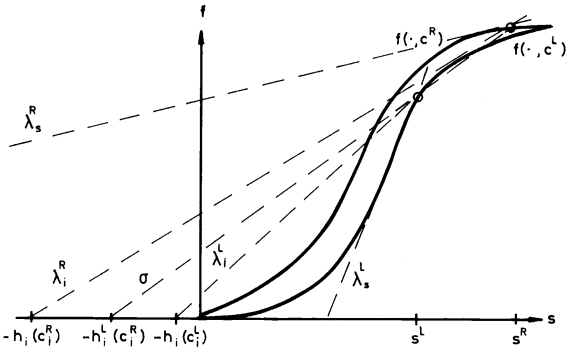


FIG. 3.3. *Overcompressive shock.*

We recall from [9] that when $n = 1$ an overcompressive shock can never be joined to another wave in a Riemann solution. As we will see below, this will not be the case when $n > 1$.

The c-shocks described above, with $c_i^L \neq c_i^R$, $c_j^L = c_j^R$ for $j \neq i$ and that satisfy (3.9), (3.10), and (3.11) will be referred to as c-shocks with multiplicity one. Shock waves where $c_i^L \neq c_i^R$, $c_{j_1}^L \neq c_{j_2}^R, \cdots, c_{j_m}^L \neq c_{j_m}^R$, and $c_k^L = c_k^R$ for $k \neq i, j_1, \cdots, j_m$ will be referred to as c-shocks with multiplicity $m+1$. In this case we obtain from (3.6) that (3.9) must hold for the index i and for any $j \in J$, where $J = \{j_1, j_2, \cdots, j_m\}$. Hence, in this case the Rankine–Hugoniot condition can be written as

$$\frac{f(s^R, c^R)}{s^R + h} = \frac{f(s^L, c^L)}{s^L + h} = \sigma, \tag{3.12}$$

where

$$h_j^L(c_j^R) = h_i^L(c_i) \equiv h \quad \text{for } j \in J. \tag{3.13}$$

We therefore observe that the structure of a shock with multiplicity $m+1$ is similar to the structure of a c-shock with multiplicity one, with the additional requirement that the relation (3.13) holds. The entropy condition for a shock with multiplicity $m+1$ is similar to (3.10) and (3.11) above, but where (3.10) is replaced by

$$c_i^L > c_i^R, \qquad c_j^L > c_j^R \quad \text{for } j \in J.$$

Finally in this section, let us consider the possible c-waves when $s = 0$ or $s = 1$. If $s = 0$ the i-rarefaction curves will simply be the lines $s = 0$, $c_j = \text{constant}$ for $j \neq i$. Furthermore, along these lines the eigenvalue λ_i is identically equal to zero. Similarly, if $s^L = s^R = 0$ the Rankine–Hugoniot condition (3.9) is satisfied with $\sigma = 0$. Hence, the

curves $s=0$, $c_j=$ constant for $j\neq i$, correspond to contact discontinuities with $\sigma=0$ and with two allowed directions. If $s=1$ the i-rarefaction curves are the lines $s=1$, $c_j=$ constant for $j\neq i$, but now the eigenvalue λ_i is given by

$$\lambda_i=\frac{1}{1+h_i(c_i)}.$$

Hence, λ_i is an increasing function of c_i along the rarefaction curves. Similarly, for given values of c^L and c^R, where $c_i^L\neq c_i^R$ and $c_j^L=c_j^R$ for $j\neq i$, the states $(1, c^L)$ and $(1, c^R)$ correspond to a c-shock with speed $1/(1+h_i^L(c_i^R))$ if and only if $c_i^L>c_i^R$. The generalization of this description to c-waves of higher multiplicity is straightforward.

4. The projection principle. A solution of a Riemann problem for the model (1.1) consists of a sequence of elementary waves that connects the left state u^L and the right state u^R. We will adopt the notation that $u^1\to u^2$ means that the left state u^1 can be connected to the right state u^2 by an elementary wave. Two elementary waves $u^1\xrightarrow{a}u^2$ and $u^2\xrightarrow{b}u^3$ 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 the final speed of the a-wave is less than or equal to the initial speed of the b-wave. Furthermore, we require a strict inequality if both waves are shock waves.

Any compatible composition of s-rarefactions and s-shocks that corresponds to a solution of the Buckley–Leverett equation (3.1) with $f=f(\cdot, c)$ for some $c\in I^n$ is referred to as an *s-wave*. We recall from the theory of the Buckley–Leverett equation that for a given left state $u^1=(s^1, c^1)$ and a given right state $u^2=(s^2, c^2)$, where $c^1=c^2=c$, there always exists a unique s-wave that connects u^1 and u^2. Furthermore, this s-wave can be constructed from a lower convex or an upper concave envelope of the function $f=f(\cdot, c)$ (cf. [9] and references given therein). This s-wave will simply be denoted by $u^1\xrightarrow{s}u^2$. A c-wave is either a c-rarefaction or a c-shock. A c-wave that connects the left state u^1 with the right state u^2 is denoted by $u^1\xrightarrow{c}u^2$.

Associated with the two-phase problem (1.1) we consider the one-phase problem given by

$$(c_i+a_i(c_i))_t+(c_i)_x=0,\qquad i=1, 2, \cdots, n. \tag{4.1}$$

This model, which is an $n\times n$ system of conservation laws, describes a one-phase flow including the same n components as the model (1.1). We observe that the n equations of (4.1) are completely decoupled. Therefore, the rarefaction waves and the shock waves of (4.1) are determined by n scalar equations. In particular, the rarefaction waves are of the form

$$c_i^L<c_i^R,\qquad \xi\equiv\frac{x}{t}=\frac{1}{1+h_i(c_i)} \tag{4.2}$$

and the shock waves of the form

$$c_i^L>c_i^R,\qquad \xi\equiv\frac{x}{t}=\frac{1}{1+h_i^L(c_i^R)}. \tag{4.3}$$

The system (4.1) is in general not strictly hyperbolic, since it might occur that $h_i(c_i)=h_j(c_j)$ for $i\neq j$. Also observe that, for any c-wave of the system (1.1), we obtain an elementary wave of (4.1) by projecting the wave curve into c-space and by letting the speed $\xi=x/t$ be given by (4.2) or (4.3). Hence, if

$$(s^1, c^1)\xrightarrow{c}(s^2, c^2)$$

is a c-wave for (1.1), we obtain a projected elementary wave for the associated system (4.1). This wave will be denoted by $c^1 \xrightarrow{P(c)} c^2$.

The following lemma explains the importance of the system (4.1) for the construction of the solution of the Riemann problem for the system (1.1).

LEMMA 4.1. *Assume that the three waves*

$$(s^L, c^L) \xrightarrow{c_1} (s^1, c) \xrightarrow{s} (s^2, c) \xrightarrow{c_2} (s^R, c^R)$$

are compatible for the system (1.1). *Then the two waves*

$$c^L \xrightarrow{P(c_1)} c \xrightarrow{P(c_2)} c^R$$

are compatible for the corresponding system (4.1).

Proof. Let $h_1, h_2 > 0$ be such that the final speed of the c_1-wave, v_1^f, and the initial speed of the c_2-wave , v_2^i, are given by

$$v_1^f = \frac{f(s^1, c)}{s^1 + h_1} \quad \text{and} \quad v_2^i = \frac{f(s^2, c)}{s^2 + h_2}.$$

By exactly the same argument as was used in the proof of Lemma 5.1 of [9], we then deduce that $h_1 \geqq h_2$ or

$$\frac{1}{1+h_1} \leqq \frac{1}{1+h_2}.$$

Hence, the final speed of $P(c_1)$ is less than or equal to the initial speed of $P(c_2)$. □

A consequence of the lemma above is that, if a solution of the Riemann problem for (1.1) with left state (s^L, c^L) and right state (s^R, c^R) is given, then the projection of the wave curves into c-space corresponds to a solution of the Riemann problem for the associated system (4.1) with data c^L and c^R. Furthermore, since (4.1) is a decoupled system, the Riemann problem for (4.1) has a unique solution for any values of c^L and c^R in $\mathbb{I}^n$. In each component the solution either consists of a single rarefaction wave or a single shock wave. The solution for the complete system is therefore obtained by superimposing the x, t-diagrams corresponding to each component (cf. Fig. 4.1). In particular we observe that the wave cones corresponding to the different components may overlap.

We have therefore seen that for any given Riemann problem for (1.1), the only possible projection of the wave curves of the solution into c-space can be determined by solving n scalar Riemann problems (cf. Figs. 4.1 and 4.2).

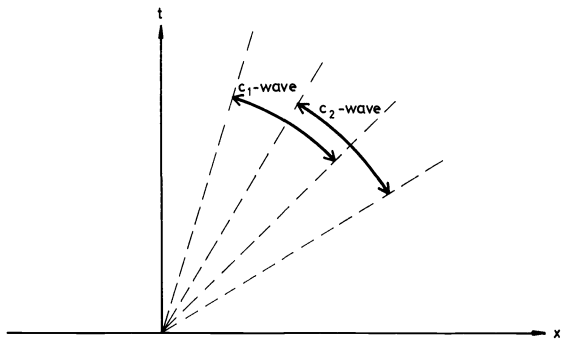


FIG. 4.1. *Component wave cones.*

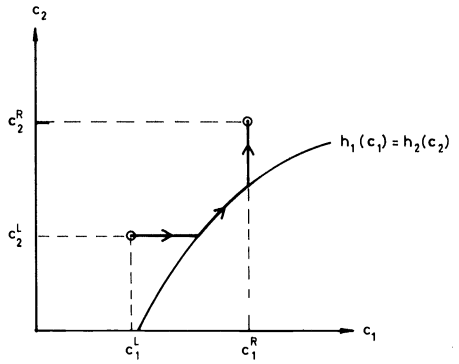


FIG. 4.2. *Projection of wave curves.*

In our construction of the solution of the Riemann problem for (1.1) we will utilize the fact that the projection of the wave curves of the solution into c-space can be easily determined. Our strategy will be first to compute the solution of the associated Riemann problem for (4.1) and thereafter the complete Riemann solution for (1.1) will be constructed from the computed sequence of projected wave curves.

5. The solution of the Riemann problem. The purpose of this section is to present the construction of the solution of the Riemann problem for the model (1.1) for arbitrary left and right states in $\mathbb{I}^{n+1}$. As above, we let $u^L = (s^L, c^L)$ and $u^R = (s^R, c^R)$ denote the left and right states, respectively. We already know from Lemma 4.1 that if a solution of the Riemann problem for (1.1) exists, then the projection of the wave curves into c-space must correspond to the wave curves of the unique solution of a Riemann problem for the corresponding model (4.1). Hence, the projection of the wave curves into c-space will form a path in c-space. This path will be referred to as the *composition path* or simply the *c-path.*

Since the c-path can easily be derived by solving n scalar Riemann problems, we will assume throughout this section that this path is given. In particular this means that c^L and c^R are fixed and that the data of the Riemann problem only varies with s^L and s^R.

In most of the discussion below we will also assume that the wave cones in x, t-space, corresponding to the Riemann solutions of the different components of (4.1), do not overlap. Hence, if the c-components are ordered with respect to increasing wave speed, we assume that

$$v_1^i \leqq v_1^f \leqq v_2^i \leqq \cdots \leqq v_n^i \leqq v_n^f, \tag{5.1}$$

where v_j^i and v_j^f denote the initial and final speed of the wave which solves the Riemann problem for the jth component of (4.1) (cf. Fig. 5.1). The assumption will be removed at the end of this section.

A consequence of (5.1) is that no c-wave occurring in the Riemann solution of (1.1) has multiplicity greater than one, i.e., no $(i, j_1, \cdots, j_m)$ wave with $m > 0$ is included in the solution. In particular, the c-path consists of n straight lines parallel to the axis. The solution of the Riemann problem for (1.1) will therefore be located in a region of (s, c)-space which can be considered as n strips in $\mathbb{R}^2$ as illustrated in Fig. 5.2, where the line $c_i = c_i^R$ is identified with the line $c_{i+1} = c_{i+1}^L$. This region will be referred to as the state space associated to the given c-path. Each state (s, c_i) in the ith strip of this region can be joined to other states by either an s-wave (i.e., $c_i = \text{const.}$) or a

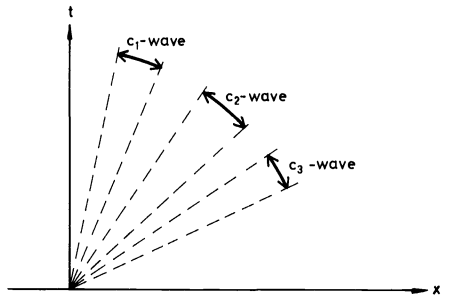


FIG. 5.1. $n=3$.

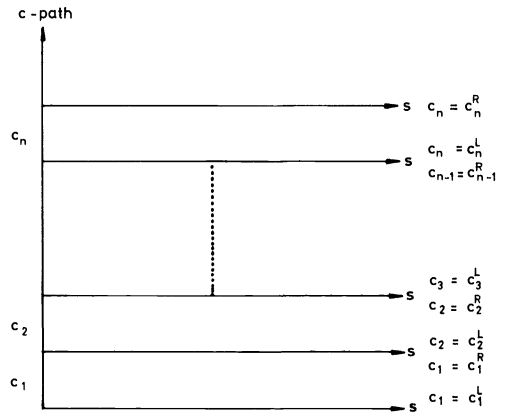


FIG. 5.2. *The state space associated with c-path.*

c-wave. The admissible c-waves are all directed upwards on Fig. 5.2. The c-waves in the ith strip are i-rarefaction if $c_i^L < c_i^R$ and i-shocks if $c_i^L > c_i^R$.

For each strip of the state space we can associate a transition curve T_i. First, for any c_i between c_i^L and c_i^R we define the associated wave speed by

$$\sigma_i(s, c_i) = \begin{cases} \dfrac{f(s, c_i)}{s + h_i(c_i)} & \text{if } c_i^L < c_i^R, \\ \dfrac{f(s, c_i)}{s + h_i^L(c_i^R)} & \text{if } c_i^L > c_i^R, \end{cases}$$

where here and below $f(s, c_i)$ denotes the value $f(s, c_1^R, \cdots, c_{i-1}^R, c_i, c_{i+1}^L, \cdots, c_n^L)$. The transition curve T_i in the ith strip is defined from the relation

$$\sigma_i(s, c_i) = f_s(s, c_i); \tag{5.2}$$

i.e., $(s, c_i) \in T_i$ if and only if $s > 0$ and (5.2) holds. For each value of c_i there is at most one value of s, called $s_i^T(c_i)$, such that $(s_i^T, c_i) \in T_i$. If no such value of s exists, we let $s_i^T(c_i) = +\infty$. We note that if $c_i^L < c_i^R$, i.e., the i-waves are rarefaction waves, the transition curve T_i corresponds exactly to the curve where $\lambda_i = \lambda_s$ (cf. § 2). On the lines $c_i = c_i^R$ or $c_{i+1} = c_{i+1}^L$ there are two possible values of s^T, $s_i^T(c_i^R)$ and $s_{i+1}^T(c_{i+1}^L)$, corresponding to the strip below and above this line. Assumption (5.1) implies that

$$s_i^T(c_i^R) \geqq s_{i+1}^T(c_{i+1}^L). \tag{5.3}$$

The transition curves T_i, in the n strips of the state space associated to the given c-path, are therefore located as illustrated in Fig. 5.3.

To construct the solution of the global Riemann problem we have to characterize the possible compatible compositions of elementary waves. In each strip of the state space this analysis corresponds to the analysis given in [9] for the case $n = 1$. To state the desired results from [9] we need some notation. For each state (s, c_i) in the ith strip of the state space we define the associated critical value $s_i^K(s, c_i)$ from the relation

$$\sigma_i(s_i^K, c_i) = \sigma_i(s, c_i). \tag{5.4}$$

If $s \geqq s_i^T(c_i)$, s_i^K is the unique value such that $s_i^K \leqq s_i^T(c_i)$ and such that (5.4) holds (cf. Fig. 5.4). If $s \leqq s_i^T(c_i)$, s_i^K is either the unique value greater than or equal to $s_i^T(c_i)$ such that (5.4) holds, or, if no such s_i^K exists, we let $s_i^K = +\infty$.

Now consider possible compositions of pairs of waves of the forms

$$(s^1, c_i^1) \xrightarrow{c} (s^M, c_i^2) \xrightarrow{s} (s^2, c_i^2) \tag{5.5}$$

and

$$(s^1, c_i^1) \xrightarrow{s} (s^M, c_i^1) \xrightarrow{c} (s^2, c_i^2) \tag{5.6}$$

in the ith strip of state space. Necessary and sufficient conditions for the compatibility of these compositions can be obtained directly from the analysis given in [9]. However,

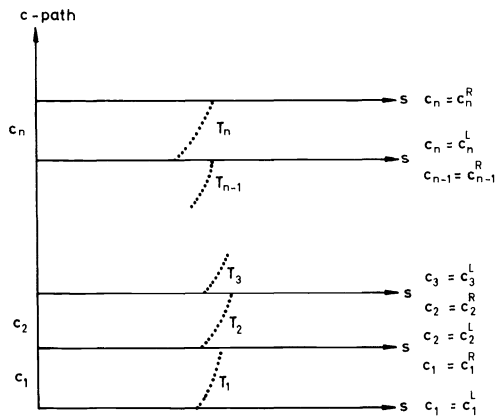


FIG. 5.3

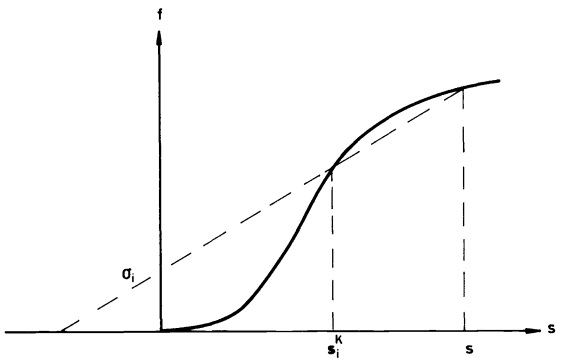


FIG. 5.4

to make these conditions as simple as possible, we will make a minor change in terminology with respect to the overcompressive c-shocks.

The admissible overcompressive shocks have the property that

$$\lambda_i^L > \sigma > \lambda_i^R \quad \text{and} \quad \lambda_s^L \geqq \sigma > \lambda_s^R,$$

where λ^L and λ^R denote eigenvalues to the left and to the right of the shock, respectively. They are illustrated in Fig. 3.3. These shocks may be thought of as a composition of a c-shock and an s-shock with the same speed. Usually such compositions are not allowed. However, in the analysis below, statements about the compatibility of compositions of elementary waves will simplify if we do allow compositions of two shocks with the same speed starting with a c-shock and do not allow the overcompressive c-shocks. With this change in terminology the overcompressive c-shocks are described as a composition of the form (5.5). Furthermore, the entropy condition for the allowed c-shock takes the form

$$\lambda_i^L > \sigma > \lambda_i^R \quad \text{and} \quad \lambda_s^L, \lambda_s^R \geqq \sigma \quad \text{or} \quad \lambda_s^L, \lambda_s^R < \sigma.$$

This change of terminology is used in all the lemmas below. The following result is now a direct consequence of Lemmas 6.1 and 7.1 of [9].

LEMMA 5.1. *Consider a composition of a c-wave and an s-wave in the ith strip of state space.*

(i) *The composition* (5.5) *is compatible if and only if*

$$s^M \leqq s_i^T(c_i^2) \quad \textit{and} \quad 0 \leqq s^2 \leqq s_i^K(s^M, c_i^2).$$

(ii) *The composition* (5.6) *is compatible if and only if*

$$s^M \geqq s_i^T(c_i^1) \quad \textit{and} \quad s_i^K(s^M, c_i^1) \leqq s^1 \leqq 1.$$

Remark. If we had not changed the terminology as described above, the statements of the lemma above would only be correct if the admissible c-waves in the ith strip were rarefaction waves. If the c-wave is a shock the proper condition for composition (5.5) is

$$s^M \leqq s_i^T(c_i^2) \quad \text{and} \quad 0 \leqq s^2 < s_i^K(s^M, c_i^2),$$

while the case $s^2 = s_i^K(s^M, c_i^2)$ corresponds to an overcompressive shock from (s^1, c_i^1) to (s^2, c_i^2). Similarly, the proper condition for (5.6) is

$$s^M \geqq s_i^T(c_i^1) \quad \text{and} \quad s_i^K(s^M, c_i^1) < s^1 \leqq 1.$$

We also note that when the c-wave is a shock, the compositions (5.5) and (5.6) only occur when $c_i^1 = c_i^L$ and $c_i^2 = c_i^R$.

Consider next the composition of three waves of the form

$$\xrightarrow{c} (s^1, c_i^M) \xrightarrow{s} (s^2, c_i^M) \xrightarrow{c}$$

on the ith strip, where we assume that $c_i^L < c_i^R$ such that the c-waves are rarefactions. If $c_i^L < c_i^M < c_i^R$ it follows from Lemma 5.1 that this composition is compatible if and only if

$$s^1 \leqq s_i^T(c_i^M), \quad s^2 \geqq s_i^T(c_i^M), \quad s^2 = s_i^K(s^1, c_i^M).$$

However, at the intersections of two strips, where the transition curve is discontinuous, there are more possibilities. Consider a composition of the form

$$\xrightarrow{c} (s^1, c_i^R) \xrightarrow{s} (s^2, c_{i+1}^L) \xrightarrow{c}. \tag{5.7}$$

The precise conditions for the compatibility of (5.7) can be derived by combining statements (i) and (ii) of Lemma 5.1. We summarize this result in the following lemma.

LEMMA 5.2. *The composition* (5.7) *is compatible if and only if either*

$$s^1 \leqq s_{i+1}^T(c_{i+1}^L) \quad \textit{and} \quad s^2 \in [s_{i+1}^K(s^1, c_{i+1}^L), s_i^K(s^1, c_i^R)]$$

or

$$s^1 \in [s_{i+1}^T(c_{i+1}^L), s_i^T(c_i^R)] \quad \textit{and} \quad s^2 \in [s_{i+1}^T(c_{i+1}^L), s_i^K(s^1, c_i^R)].$$

The two cases of the lemma are illustrated in Figs. 5.5 and 5.6.

Before we proceed to construct the solution of the Riemann problem for (1.1) we will first review the main structure of the solution in the case when $n = 1$. For a more precise description of this solution we refer to [9]. Consider first the case when $c_1^L < c_1^R$, i.e., when the c-waves are rarefactions. We divide this case further into two subcases.

First, assume that $s^L < s_1^T(c_1^L)$ and that the rarefaction curve through the point (s^L, c_1^L) intersects the line $c_1 = c_1^R$ at a point $(\hat{s}, c_1^R)$, where $\hat{s} \leqq s_1^T(c_1^R)$. Typical solutions of the Riemann problem, depending on the location of s^R, are illustrated in Fig. 5.7, where $\hat{s}^K = s_1^K(\hat{s}, c_1^R)$. In particular, the Riemann solution terminates with a c-wave if and only if $s^R = \hat{s}$ and $s^R > \hat{s}^K$.

The second subcase occurs when either $s^L > s_1^T(c_1^L)$ or $s^L \leqq s_1^T(c_1^L)$, but the rarefaction curve through (s^L, c_1^L) intersects the transition curve at $c_1 = c_1^*$, where $c_1^* < c_1^R$. Examples of Riemann solutions in this case are illustrated in Figs. 5.8(a) and 5.8(b). In particular, the solution terminates with a c-wave if and only if $s^R \geqq s_1^T(c_1^R)$. Hence, if we let $\hat{s} = \hat{s}^K = s_1^T(c_1^R)$ in this latter case, we conclude that, if the c-waves are rarefaction waves, the unique solution of the Riemann problem terminates with a c-wave if and only if $s^R = \hat{s}$ or $s^R > \hat{s}^K$.

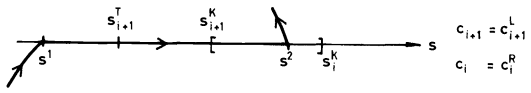


FIG. 5.5. $s^1 \leqq s_{i+1}^T(c_{i+1}^L)$.

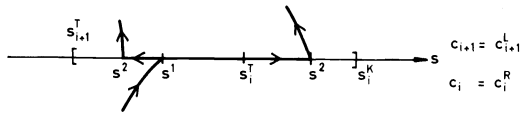


FIG. 5.6. $s_{i+1}^T < s^1 < s_i^T$.

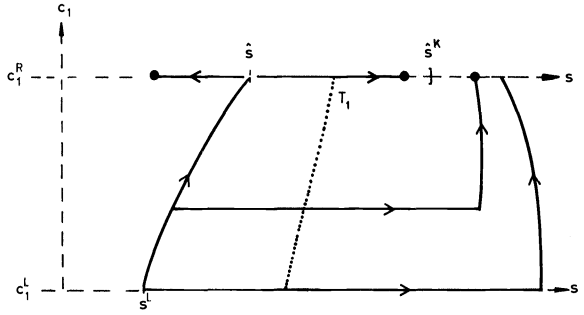


FIG. 5.7

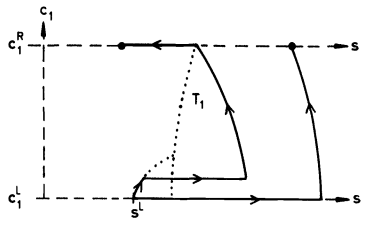


FIG. 5.8(a). $s^L \leqq s_1^T(c_1^L)$.

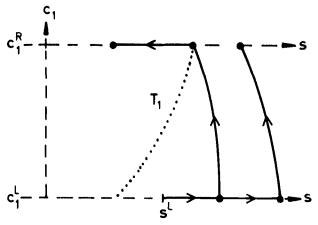


FIG. 5.8(b). $s^L > s_1^T(c_1^L)$.

Consider next the case when $n=1$ and $c_1^L > c_1^R$. Again this case is divided into two subcases. If $s^L \leqq s_1^T(c_1^L)$ there is always a unique value $\hat{s}$ such that $\hat{s} \in s_1^T(c_1^R)$ and such that

$$(s^L, c_1^L) \xrightarrow{c} (\hat{s}, c_1^R)$$

is a c-shock. Let $\hat{s}^K = s_1^K(\hat{s}, c_1^R)$. The different Riemann solutions in this case are illustrated in Fig. 5.9.

Finally, if $s^L > s_1^T(c_1^L)$ let s_1^- be the unique value $s_1^- < s_1^T(c_1^R)$ such that

$$(s_1^T(c_1^L), c_1^L) \xrightarrow{c} (s_1^-, c_1^R)$$

is a c-shock. Furthermore, let $s_1^+ = s_1^K(s_1^-, c_1^R)$. The different possible Riemann solutions are illustrated in Fig. 5.10.

Hence, if we let $\hat{s} = s_1^-$ and $\hat{s}^K = s_1^+$ in this latter case, the structure of Riemann solutions in the shock case has the property that it terminates with a c-shock if and only if $s^R = \hat{s}$ or $s^R > \hat{s}^K$. We have therefore seen that this property of the Riemann solution is shared both by the rarefaction case and the shock case.

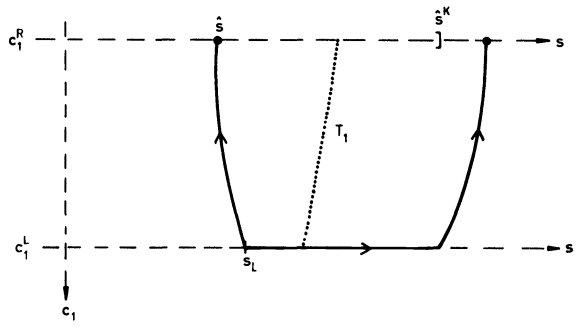


FIG. 5.9

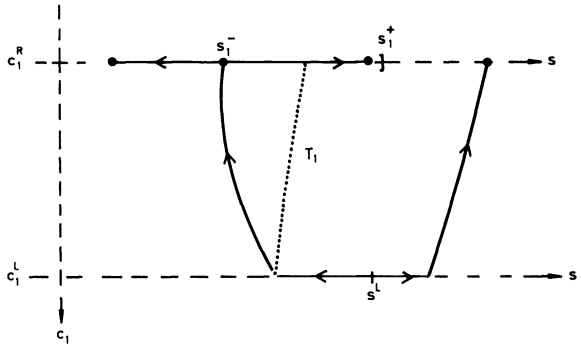


FIG. 5.10

(We remark again that the correctness of this description of the Riemann solution depends on the change of terminology introduced at the beginning of this section. A more correct description of the solution in the shock case is that it terminates with a c-shock if and only if $s^R = \hat{s}$ or $s^R \geqq \hat{s}^K$.)

Let us return to the case when $n > 1$. Motivated by the structure of the Riemann solution above, we introduce some extra notation. If the i-waves are shocks, we define s_i^- as the unique value satisfying

$$s_i^- < s_i^T(c_i^R) \text{ and such that } (s_i^T(c_i^L), c_i^L) \xrightarrow{c} (s_i^-, c_i^R)$$

corresponds to a c-shock. Furthermore, we let $s_i^+ = s_i^K(s_i^-, c_i^R)$. The location of these points on the line $c_i = c_i^R$ and $c_{i+1} = c_{i+1}^L$ are illustrated in Fig. 5.11. The two s_i^--points indicate that s_i^- can be located on any side of T_{i+1}, but always to the left of T_i. In particular,

$$s_i^+ \geqq s_i^T(c_i^R) \geqq s_{i+1}^T(c_{i+1}^L),$$

while s_i^- can be greater than or less than $s_{i+1}^T(c_{i+1}^L)$. If the i-waves are rarefactions we let $s_i^- = s_i^+ = s_i^T(c_i^R)$.

We are now in a position to construct the general solution of the Riemann problem for (1.1).

We recall again that since c^L and c^R are considered fixed, the data of the Riemann problem only varies with s^L and s^R.

LEMMA 5.3. *Assume that the associated Riemann solution of* (4.1) *satisfies condition* (5.1). *The Riemann problem for* (1.1) *has a unique solution for arbitrary* s^L *and* s^R. *Furthermore, for any given* s^L *there exist values* $\hat{s}$ *and* $\hat{s}^K$, *with* $\hat{s} \leqq s_n^-$ *and* $\hat{s}^K \geqq s_n^+$, *such that the Riemann solution terminates with c-wave if and only if*

$$s^R = \hat{s} \quad \text{or} \quad s^R > \hat{s}^K.$$

Here $\hat{s}$ *and* $\hat{s}^K$ *are related by* $\hat{s}^K = s_n^K(\hat{s}, c_n^R)$.

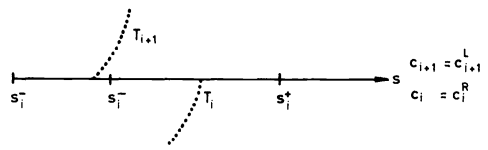


FIG. 5.11

Proof. The proof uses induction on n. We have already seen in the discussion above that the result holds for $n=1$. Assume that the result holds for $n-1$; i.e., any states (s^L, c_1^L) and (s^R, c_{n-1}^R) can be connected by a unique composition of compatible elementary waves.

Throughout the rest of the proof we consider s^L as fixed. By the induction hypothesis there exist values $\hat{s}_{n-1}$ and $\hat{s}_{n-1}^K$, where $\hat{s}_{n-1} \leqq s_{n-1}^-$ and $\hat{s}_{n-1}^K = s_{n-1}^K(\hat{s}_{n-1}, c_{n-1}^R)$, such that a Riemann solution that connects (s^L, c_2^L) to (s^R, c_{n-1}^R) terminates with a c-wave if and only if $s^R = \hat{s}_{n-1}$ or $s^R > \hat{s}_{n-1}^K$.

To show that the desired result holds for n, we have to construct the Riemann solution for any right state (s^R, c_n^R) and to show that this solution is unique. In the analysis below we will concentrate our effort on the construction of a solution. The uniqueness of this solution can easily be established in all cases by applying the results of Lemmas 5.1 and 5.2. We will start the construction by identifying the desired values $\hat{s}$ and $\hat{s}^K$.

Assume first that the n-waves are rarefactions; i.e., assume that $c_n^L < c_n^R$. If $\hat{s}_{n-1} < s_n^T(c_n^L)$ and if the rarefaction curve through $(\hat{s}_{n-1}, c_n^L)$ intersects the line $c_n = c_n^R$, we let $\hat{s}$ be the unique value less than or equal to $s_n^T(c_n^R)$ such that the rarefaction curve intersects the line $c_n = c_n^R$ at $(\hat{s}, c_n^R)$. In this case the composition

$$(s^L, c_1^L) \Rightarrow (\hat{s}_{n-1}, c_{n-1}^R) \xrightarrow{c} (\hat{s}, c_n^R),$$

where here and below $\Rightarrow$ denotes the unique Riemann solution from the induction hypothesis, is a Riemann solution that terminates with a c-wave. If $\hat{s}_{n-1} \geqq s_n^T(c_n^L)$ or if no rarefaction curve connects $(\hat{s}_{n-1}, c_n^L)$ to the line $c_n = c_n^R$, we let $\hat{s} = s_n^T(c_n^R)$. To see that the Riemann solution with $s^R = \hat{s}$ terminates with a c-wave, in this case we let s_{n-1} be the unique value such that $s_{n-1} > s_n^T(c_n^L)$ and such that (s_{n-1}, c_n^L) and $(\hat{s}, c_n^R)$ are connected by a rarefaction curve (cf. Fig. 5.12).

If $s_{n-1} > \hat{s}_{n-1}^K$ it follows from the induction hypothesis that the solution

$$(s^L, c_1^L) \Rightarrow (s_{n-1}, c_{n-1}^R)$$

terminates with a c-wave. Therefore, the composition

$$(s^L, c_1^L) \Rightarrow (s_{n-1}, c_n^L) \xrightarrow{c} (s_n, c_n^R)$$

is compatible and terminates with a c-wave. If $s_{n-1} \leqq \hat{s}_{n-1}^K$ it follows from the result for $n=1$ that the unique compatible composition that connects $(\hat{s}_{n-1}, c_n^L)$ and $(\hat{s}, c_n^R)$ either starts and terminates with a c-wave or is of the form

$$(\hat{s}_{n-1}, c_n^L) \xrightarrow{s} (s_{n-1}, c_n^L) \xrightarrow{c} (\hat{s}, c_n^R). \tag{5.8}$$

In particular, the states are connected by waves of the form (5.8) only if $\hat{s}_{n-1} \geqq s_n^T(c_n^L)$ or if $s_{n-1} \in [s_n^K(\hat{s}_{n-1}, c_n^L), \hat{s}_{n-1}^K]$. But in these cases it follows from Lemma 5.2 and the

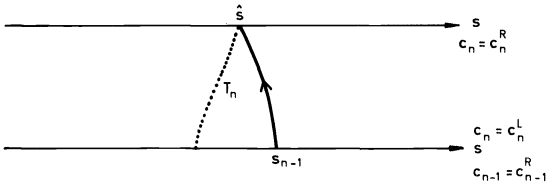


FIG. 5.12

induction hypothesis that the complete composition

$$(s^L, c_1^L) \Rightarrow (\hat{s}_{n-1}, c_n^L) \xrightarrow{s} (s_{n-1}, c_n^L) \xrightarrow{c} (\hat{s}, c_n^R)$$

is compatible (cf. Figs. 5.13 a,b). We have therefore seen that if the n-waves are rarefactions, then in all possible cases a value $\hat{s} \leqq s_n^T(c_n^R)$ can be constructed such that if $s^R = \hat{s}$ the Riemann solution terminates with a c-wave.

Consider next the case where the n-waves are shocks. If $\hat{s}_{n-1} \leqq s_n^T(c_n^L)$, we simply let $\hat{s}$ be the unique value such that $\hat{s} \leqq s_n^-$ and such that

$$(\hat{s}_{n-1}, c_n^L) \xrightarrow{c} (\hat{s}_n, c_n^R)$$

is a c-shock. If $\hat{s}_{n-1} > s_n^T(c_n^L)$, we let $\hat{s} = s_n^-$. Since it follows from Lemma 5.2 and the induction hypothesis that the composition

$$(s^L, c_1^L) \Rightarrow (\hat{s}_{n-1}, c_n^L) \xrightarrow{s} (s_n^T(c_n^L), c_n^L) \xrightarrow{c} (\hat{s}, c_n^R)$$

(cf. Fig. 5.14) is compatible, we have, in all cases where the n-waves are shocks, generated $\hat{s} \leqq s_n^-$ such that the Riemann solution with $s^R = \hat{s}$ terminates with a c-wave. Hence, we have completed the construction of the values $\hat{s}$ and $\hat{s}^K \equiv s_n^K(\hat{s}, c_n^R)$.

Furthermore, it follows from Lemma 5.1 that if the solution that connects (s^L, c_1^L) and $(\hat{s}, c_n^R)$ is extended by an s-wave of the form

$$(\hat{s}, c_n^R) \xrightarrow{s} (s^R, c_n^R),$$

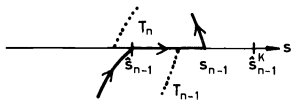


FIG. 5.13(a). $\hat{s}_{n-1} \geqq s_n^T(c_n^L)$.

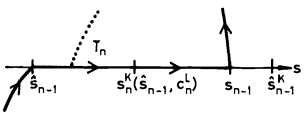


FIG. 5.13(b). $s_{n-1} \in [s_n^K(\hat{s}_{n-1}, c_n^L), \hat{s}_{n-1}^K]$.

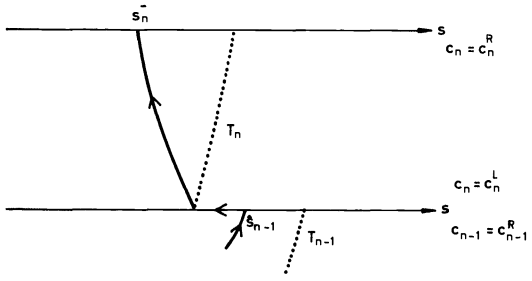


FIG. 5.14

then the complete composition is compatible as long as $s^R \leqq \hat{s}^K$. Hence, to complete the proof of the lemma, we have to construct a Riemann solution that terminates with a c-wave when $s^R > \hat{s}^K$. If $s^R > \hat{s}^K \geqq s_n^+$ there exists a unique value $s_{n-1} > s_n^T(c_n^L)$ such that

$$(s_{n-1}, c_n^L) \xrightarrow{c} (s^R, c_n^R)$$

is a c-wave (shock or rarefaction). If $s_{n-1} > \hat{s}_{n-1}^K$ the induction hypothesis implies that the Riemann solution is given by

$$(s^L, c_1^L) \Rightarrow (s_{n-1}, c_n^L) \xrightarrow{c} (s^R, c_n^R).$$

Otherwise, if $s_{n-1} \leqq s_{n-1}^K$ we first consider the compatible composition that connects $(\hat{s}_{n-1}, c_n^L)$ and (s^R, c_n^R). This composition either starts and terminates with a c-wave (this is only possible if $\hat{s}_{n-1} < s_n^T(c_n^L)$ and the n-waves are rarefactions) or it is of the form

$$(\hat{s}_{n-1}, c_n^L) \xrightarrow{s} (s_{n-1}, c_n^L) \xrightarrow{c} (s^R, c_n^R). \tag{5.9}$$

As above, the composition is of the form (5.9) only if $\hat{s}_{n-1} \geqq s_n^T(c_n^L)$ or if $s_{n-1} \in [s_n^K(\hat{s}_{n-1}, c_n^L), \hat{s}_{n-1}^K]$. Hence, Lemma 5.2 and the induction hypothesis imply that

$$(s^L, c_1^L) \Rightarrow (\hat{s}_{n-1}, c_n^L) \xrightarrow{s} (s_{n-1}, c_n^L) \xrightarrow{c} (s^R, c_n^R)$$

is a compatible composition of waves. We have therefore seen that the Riemann solution always terminates with a c-wave when $s^R > \hat{s}^K$. This completes the induction argument and therefore the proof of the lemma. □

The construction of the Riemann solution given above can be thought of as a "factorization" of the general Riemann problem associated with a given c-path. First, the values $\hat{s}_i$ and $\hat{s}_i^K$, for $i = 1, 2, \cdots, n$, are constructed from s^L and the given c-path. Thereafter, the Riemann solution is constructed from s^R and all the values $\hat{s}_i$ and $\hat{s}_i^K$. This interpretation of the proof makes a computer implementation of the construction rather attractive.

The proof of Lemma 5.3 above completes the discussion of the Riemann problem for (1.1) under the additional assumption (5.1). To remove this assumption, we finally do allow the wave cones, corresponding to the different components of the system (4.1), to overlap. Hence, c-waves of multiplicity greater than one may occur in the Riemann solution of (1.1). However, since the structure of such c-waves is similar to the structure of simple waves, their occurrence will not change the logical structure of the construction of the Riemann solution. The occurrence of c-waves of multiplicity greater than one will only increase the possible number of strips of the state space along the given c-path. For example, if $n = 2$ and the two wave cones intersect, the state space along the c-path will consist of three strips as illustrated in Figs. 5.15 and 5.16. Hence, by possibly increasing the number of strips to at most $2n - 1$, the construction given in the proof of Lemma 5.3 also applies when assumption (5.1) is removed. We have therefore established the main result of this paper.

THEOREM 5.1. *For arbitrary states* $u^L = (s^L, c^L) \in \mathbb{I}^{n+1}$ *and* $u^R = (s^R, c^R) \in \mathbb{I}^{n+1}$ *there exists a unique solution of the Riemann problem for the system* (1.1) *with left state* u^L *and right state* u^R.

Before we end this section we would like to point out one property of the Riemann solution constructed above. We recall from [9] that in the case when $n = 1$ an overcompressive shock can never join any other wave in a Riemann solution. However, it can

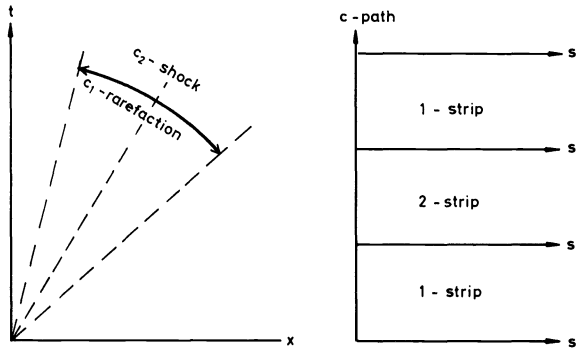


FIG. 5.15

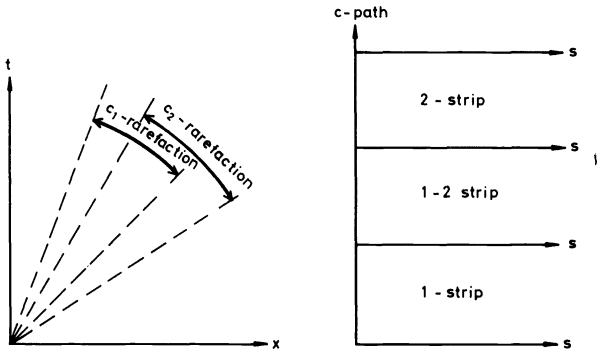


FIG. 5.16

be seen from the construction above that this property does not hold when $n > 1$. An example of a Riemann solution of the form

$$(s^L, c_1^L, c_2^L) \xrightarrow{c} (\bar{s}, c_1^R, c_2^L) \xrightarrow{c} (s^R, c_1^R, c_2^R),$$

where both c-waves are shocks and where the first shock is overcompressive, is illustrated in Fig. 5.17.

6. Conclusion. The solution of the general $(n+1)\times(n+1)$ Riemann problem for the model (1.1) has been constructed. The construction is based on the fact that a projection principle, described in § 4, decouples the Riemann problem into a finite sequence of coupled 2×2 Riemann problems. This sequence of coupled Riemann problems is then solved in § 5. The construction done in § 5 establishes a factorization algorithm for the coupled sequence of Riemann problems. In a forthcoming paper we will discuss further the practical implementation of this algorithm and also present examples of numerical results.

The projection principle described in § 4 is valid independent of the structure of the adsorption functions a_i. Hence, if for example the functions $a_i(c_i)$ in (1.1) are replaced by a family of generalized Langmuir adsorption functions of the form (1.2), the general Riemann solution can still be found from a construction of the form studied in § 5. The only difference would be that the corresponding one-phase problem would be changed. Hence, a different procedure for the construction of the desired composition path would be required. In a forthcoming paper we will discuss how the techniques

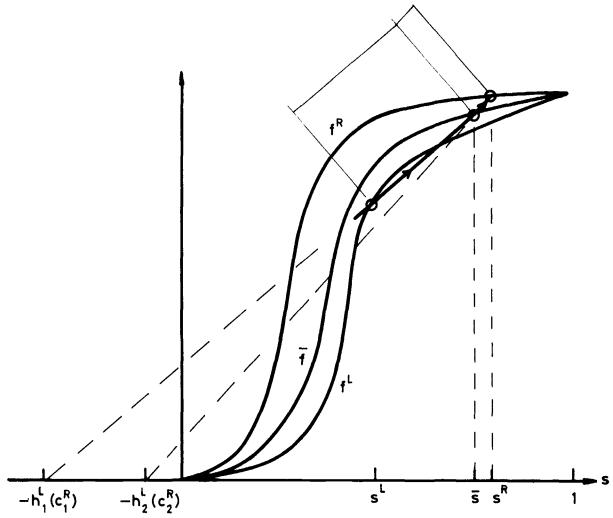


FIG. 5.17. $f^L = f(\cdot, c_1^L, c_2^L)$, $\bar{f} = f(\cdot, c_1^R, c_2^L)$, $f^R = f(\cdot, c_1^R, c_2^R)$.

developed in this paper can be combined with the results of [13] to obtain the Riemann solution of a two-phase model of the form (1.1), but where the functions $a_i(c_i)$ are replaced by a family of generalized Langmuir adsorption functions.

REFERENCES

[1] A. I. CHORIN, *Random choice solutions of hyperbolic systems*, J. Comput. Phys., 22 (1976), pp. 517-533.

[2] P. CONSUS AND W. PROSKUROWSKI, *Numerical solutions of a nonlinear hyperbolic equation by the random choice method*, J. Comp. Phys., 30 (1979), pp. 153-166.

[3] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems*, Comm. Pure Appl. Math., 18 (1965), pp. 697-715.

[4] J. GLIMM, E. ISAACSON, D. MARCHESIN, AND O. MCBRYAN, *Front tracking for hyperbolic systems*, Adv. in Appl. Math., 2 (1981), pp. 91-119.

[5] J. GLIMM, B. LINDQUIST, O.MCBRYAN, AND L. PADMANABHAN, *A front tracking reservoir simulator I: the water coning problem*, Frontiers Appl. Math., 1 (1983).

[6] S. GODUNOV, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb. (N.S.), 47 (1959), pp. 271-290.

[7] E. ISAACSON, *Global solution of a Riemann problem for a non-strictly hyperbolic system of conservation laws arising in enhanced oil recovery*, J. Comp. Phys., to appear.

[8] E. ISAACSON AND B. TEMPLE, *Analysis of a singular hyperbolic system of conservation laws*, J. Differential Equations, 65 (1986), pp. 250-268.

[9] T. JOHANSEN AND R. WINTHER, *The solution of the Riemann problem for a hyperbolic system of conservation laws modelling polymer flooding*, SIAM J. Math. Anal., 19 (1988), pp. 541-566.

[10] 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.

[11] L. W. LAKE AND F. HELFFERICH, *Cation exchange in chemical flooding: part 2—the effect of dispersion, cation exchange, and polymer/surfactant adsorption on chemical flood environment*, Soc. Pet. Enrgr. J., (Dec. 1978), pp. 435-444.

[12] P. D. LAX, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537-566.

[13] H. RHEE, R. ARIS, AND N. R. AMUNDSON, *On the theory of multicomponent chromatography*, Philos. Trans. Roy. Soc. London Ser. A, 267 (1970), pp. 250-268.

[14] D. G. SCHAEFFER AND M. SCHEARER, *Riemann problems for nonstrictly hyperbolic* 2×2 *systems of conservation laws*, Trans. Amer. Math. Soc. 304 (1987), pp. 267-306.

[15] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York 1982.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 930-938, July 1989

009

PERSISTENCE IN DISCRETE SEMIDYNAMICAL SYSTEMS*

H. I. FREEDMAN† AND JOSEPH W.-H. SO‡

Abstract. Conditions are established for the persistence of a discrete semidynamical system formulated in terms of the global attractor of the boundary semiflow and its stable set. An application to an ecological model of a predator-prey system is given.

Key words. semidynamical systems, persistence, dissipative, isolated invariant set, stable and unstable sets, acyclic covering, predator-prey

AMS(MOS) subject classification. 92A17

1. Introduction. Let X be a metric space with metric d. A map $f: X \to X$ defines a discrete semidynamical system $T: \mathbb{Z}_+ \times X \to X$ by $T(n, x) = f^n(x)$, where $\mathbb{Z}_+$ denotes the set of nonnegative integers and $f^n(x)$ denotes the nth iterate of x under f, i.e., $f^n(x) = f \circ f \circ \cdots \circ f(x)$ (n times). Such discrete semidynamical systems are frequently used in the modeling of ecological systems where X is the set of all possible states of the populations in the ecosystem (for example, $X = \mathbb{R}^n_+$, the nonnegative cone in $\mathbb{R}^n$) and f describes what happens to each state after a fixed period of time (for example, a year or a generation).

Let Y be a subspace of X. We say that f is persistent (with respect to Y) if for all $x \in X \setminus Y$, $\liminf_{n\to\infty} d(f^n(x), Y) > 0$. This means that Y is in some sense a repeller (or is ejective). Stronger or weaker forms than the above notion of persistence have also been introduced (Butler, Freedman, and Waltman [1]). In the context of ecological modeling, Y could be the set of extinction states in X (for example, $Y = \partial(\mathbb{R}^n_+)$, the boundary of $\mathbb{R}^n_+$ in $\mathbb{R}^n$). In that case, persistence captures the idea of nonextinction or coexistence.

The object of this paper is to obtain criteria for persistence that are testable at least in some elementary applications. Roughly speaking, these criteria are conditions imposed on the global attractor of $f|_Y$, the restriction of f on Y. More specifically, we will derive the discrete semidynamical systems analogue of the results given in Butler and Waltman [2].

Recently, Fonda [4] has derived criteria for persistence in discrete semidynamical systems of ecological models. His criteria involve establishing the existence of a certain persistence function. However, the question of how to construct such a persistence function for a given model remains open. We emphasize that our technique is a testable one in that the criteria can be checked provided only that the boundary dynamics can be analyzed.

This paper utilizes the notation and theory of dynamical systems (see [13]), as modified to discrete semidynamical systems (see [12]). The required modified definitions are stated in § 2 for completeness.

* Received by the editors January 27, 1988; accepted for publication (in revised form) July 1, 1988.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant NSERC A4823.

‡ Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322. Present address, Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant NSERC OGP36475 and the University of Alberta Central Research Fund.

The rest of the paper is organized as follows. In § 2, we introduce some basic notation and definitions in the theory of discrete semidynamical systems. We prove in Propositions 2.2 and 3.2 that an acyclic covering gives rise to a Morse decomposition. Our main results on persistence of discrete semidynamical systems (Theorem 3.3) are presented in § 3. We also prove a Butler-McGehee type lemma for discrete semidynamical systems (Theorem 3.1) in this section. Section 4 consists of an application of this theory to an ecological model of a predator-prey system. Finally, a brief discussion and concluding remarks are given in § 5.

2. Preliminaries. Let $Y \subset X$ and let $f: X \to X$. Let $f^n(x)$ denote the nth iterate of x under f.

DEFINITION 2.1. f is said to be *persistent* (with respect to Y) if for all $x \in X \setminus Y$, $\liminf_{n\to\infty} d(f^n(x), Y) > 0$.

The main objective of this paper is to answer the following question.

PERSISTENCE QUESTION. When do we have persistence (with respect to Y)?

In [2] Butler and Waltman study the Persistence Question for a continuous dynamical system. Our aim here is to carry out the same program as in the above paper for the discrete semidynamical system defined by f. As we shall see, the results we obtain are similar. However, the noninvertibility of f makes the proofs more difficult.

In the applications we have in mind $X = \mathbb{R}^n_+$, the nonnegative cone in $\mathbb{R}^n$. It is the set of all possible states of the populations in an ecosystem. $Y = \partial(\mathbb{R}^n_+)$, the boundary of $\mathbb{R}^n_+$ in $\mathbb{R}^n$, is the set of "extinction" states of the populations in the ecosystem. f is a map that describes what happens to a state of populations after a certain fixed period of time, such as after one year or one generation. Note that f is autonomous, in the sense that its action is the same independent of year or generation.

Based on the above discussion, we will make the following assumptions.

(A1) X is a metric space with metric d.

(A2) Y is a closed subset of X.

(A3) $f: X \to X$ is continuous.

(A4) $f(Y) \subset Y$.

(A5) $f(X \setminus Y) \subset X \setminus Y$.

Under these assumptions, $f|_Y : Y \to Y$ is continuous.

DEFINITION 2.2. Let $x \in X$. Let $\mathbb{Z}$ denote the set of integers and let $\mathbb{Z}_+$ denote the set of nonnegative integers.

(i) A sequence $\{x_n\}_{n\in\mathbb{Z}_+}$ of points in X is called a *positive orbit* through x if $x_0 = x$ and $f(x_n) = x_{n+1}$ for all $n \in \mathbb{Z}_+$.

(ii) A sequence $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ of points in X is called a *negative orbit* through x if $x_0 = x$ and $f(x_{-n-1}) = x_{-n}$ for all $n \in \mathbb{Z}_+$.

(iii) A sequence $\{x_n\}_{n\in\mathbb{Z}}$ of points in X is called an *orbit* through x if $x_0 = x$ and $f(x_n) = x_{n+1}$ for all $n \in \mathbb{Z}$.

The positive orbit through x always exists and is unique. In fact, it is the sequence $O^+(x) = \{x, f(x), f^2(x), \cdots\}$. The negative orbit and the orbit through x may not exist, and even when they exist, they may not be unique.

DEFINITION 2.3. A positive orbit $\{x_n\}_{n\in\mathbb{Z}_+}$ (respectively, a negative orbit $\{x_{-n}\}_{n\in\mathbb{Z}_+}$) through x is said to be *compact* if, when considered as a subset of X, it is precompact.

(A6) For all $x \in X$, $O^+(x)$ is a compact positive orbit.

DEFINITION 2.4. (i) Let $\{x_n\}_{n\in\mathbb{Z}_+}$ be a positive orbit. The *omega limit set* of $\{x_n\}_{n\in\mathbb{Z}_+}$ is the set $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) = \{y \in X : y = \lim_{n\to\infty} x_{i_n}$ for some subsequence $\{x_{i_n}\}_{n\in\mathbb{Z}_+}$ of $\{x_n\}_{n\in\mathbb{Z}_+}\}$.

(ii) Let $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ be a negative orbit. The *alpha limit set* of $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ is the set $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) = \{y \in X : y = \lim_{n\to\infty} x_{-i_n}$ for some subsequence $\{x_{-i_n}\}_{n\in\mathbb{Z}_+}$ of $\{x_{-n}\}_{n\in\mathbb{Z}_+}\}$.

The omega limit set (respectively, alpha limit set) of a positive orbit (respectively negative orbit) is the set of all limit points of the positive orbit (respectively, negative orbit) when considered as a sequence. We also denote the omega limit set of the positive orbit through x by $\Lambda^+(x)$.

DEFINITION 2.5. Let $M \subset X$. M is *positively invariant* (respectively, *negatively invariant, invariant*) if $f(M) \subset M$ (respectively, $M \subset f(M)$, $f(M) = M$).

The union of invariant sets is invariant. For a continuous f, the closure of a positively invariant set is positively invariant, the closure of a precompact negative invariant set is negatively invariant, and the closure of a precompact invariant set is invariant.

DEFINITION 2.6. Let $M \subset X$ be invariant. M is said to be *compactly invariantly connected* if whenever $M \subset M_1 \cup M_2$, where M_1 and M_2 are disjoint nonempty, compact, invariant sets, then either $M \cap M_1 = \emptyset$ or $M \cap M_2 = \emptyset$.

PROPOSITION 2.1. (i) *If $\{x_n\}_{n\in\mathbb{Z}_+}$ is a compact positive orbit, then $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+})$ is nonempty, compact, invariant, and compactly invariantly connected, and* $\lim_{k\to\infty} d(x_k, \Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+})) = 0$.

(ii) *If $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ is a compact negative orbit, then $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is nonempty, compact, invariant, and compactly invariantly connected, and* $\lim_{k\to\infty} d(x_{-k}, \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})) = 0$.

Proof. (i) See Theorem 1.5.2 of [12].

(ii) Since $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ is a sequence in a compact set, it has a convergent subsequence. Therefore, $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is nonempty. Since the set of all limit points of a sequence is closed, therefore $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is closed and hence it is compact. Let $y \in \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$. Then $y = \lim_{n\to\infty} x_{-i_n}$ for some subsequence $\{x_{-i_n}\}_{n\in\mathbb{Z}_+}$ of $\{x_{-n}\}_{n\in\mathbb{Z}_+}$. Clearly, $f(y) = \lim_{n\to\infty} x_{-i_n+1} \in \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$. On the other hand, if z is the limit of any convergent subsequence of $\{x_{-i_n-1}\}_{n\in\mathbb{Z}_+}$ then $z \in \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ and $f(z) = y$. Thus, $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is invariant.

To show that $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is compactly invariantly connected, let M_1 and M_2 be two disjoint nonempty, compact, invariant sets such that $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \subset M_1 \cup M_2$. Then $\varepsilon = d(M_1, M_2) > 0$. Let $U_i = B_{\varepsilon/3}(M_i)$, the $\varepsilon/3$-ball about M_i. Then $M_i \subset U_i$ $(i = 1, 2)$ and $U_1 \cap U_2 = \emptyset$. For all $x \in M_1, f(x) \in M_1 \subset U_1$. Therefore, there exists $\delta_x > 0$ such that $\mathrm{cl}\,(B_{\delta_x}(x)) \subset U_1$ and $f(\mathrm{cl}\,(B_{\delta_x}(x))) \subset U_1$. Since M_1 is compact, there exist $x_1, \cdots, x_K \in M_1$ such that $M_1 \subset V_1$, $\mathrm{cl}\,(V_1) \subset U_1$, and $f(\mathrm{cl}\,(V_1)) \subset U_1$, where $V_1 = \cup_{i=1}^K B_{\delta_{x_i}}(x_i)$. Suppose that $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \cap M_i \neq \emptyset$ for $i = 1, 2$. Then $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ intersects both V_1 and U_2 an infinite number of times. This implies the existence of subsequences $\{x_{-1_n}\}_{n\in\mathbb{Z}_+}$, $\{x_{-2_n}\}_{n\in\mathbb{Z}_+}$ of $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ such that $x_{-1_n} \in V_1$ and $x_{-2_n} \in U_2$. Define the subsequence $\{x_{-3_n}\}_{n\in\mathbb{Z}_+}$ of $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ by the following properties: $-1_n \leqq -3_n < -2_n$, $x_k \in V_1$ for all k, $-1_n \leqq k \leqq -3_n$ and $x_{-3_n+1} \notin V_1$ for all $n \in \mathbb{Z}_+$. Since $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is compact, by going to subsequences if necessary, we can assume that $\{x_{-3_n}\}_{n\in\mathbb{Z}_+}$ converges. Let $p = \lim_{n\to\infty} x_{-3_n+1} = \lim_{n\to\infty} f(x_{-3_n})$. Since $x_{-3_n+1} \in U_1 \setminus V_1$ for all $n \in \mathbb{Z}_+$, $p \in \bar{U}_1 \setminus V_1$. Thus, $p \in \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is neither in V_1 nor U_2. This contradicts $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \subset V_1 \cup U_2$.

Suppose $\limsup_{k\to\infty} d(x_{-k}, \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})) > 0$. Then there exists $\eta > 0$ and a subsequence $\{x_{-i_k}\}_{k\in\mathbb{Z}_+}$ of $\{x_{-k}\}_{k\in\mathbb{Z}_+}$ such that $d(x_{-i_k}, \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})) \geqq \eta$. Since $\{x_{-i_k}\}_{k\in\mathbb{Z}_+}$ contains a convergent subsequence whose limit is in $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$, this is a contradiction. □

Under assumptions (A2), (A3), and (A6), f is not persistent if and only if there exists $x \in X \setminus Y$ such that $\Lambda^+(x) \cap Y \neq \emptyset$ (that is, $d(\Lambda^+(x), Y) = 0$).

DEFINITION 2.7. f is said to be dissipative if the set $\Omega(f) = \cup \{\Lambda^+(x): x \in X\}$ is precompact.

Under assumptions (A3) and (A6), if f is dissipative then $\Omega(f)$ is invariant and so is its closure, $\bar{\Omega}(f)$.

DEFINITION 2.8. A nonempty, closed invariant subset M of X is an *isolated invariant set* if it is the maximal (under the order of inclusion) invariant set in some neighbourhood of itself. A subset N of X is an *isolating neighbourhood* if the maximal invariant set in N is nonempty, closed, and contained in int (N), the interior of N.

Let M be an isolated invariant set and let $x \in X \setminus M$. Then there exists a closed isolating neighbourhood N of M such that $x \in X \setminus N$.

DEFINITION 2.9. Let $M \subset X$ be an isolated invariant set.

(i) A compact positive orbit $\{x_n\}_{n\in\mathbb{Z}_+}$ is said to be in the *stable set* of M (under f) (in notation, $\{x_n\}_{n\in\mathbb{Z}_+} \in W^+(M)$) if $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) \subset M$.

(ii) A compact negative orbit $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ is said to be in the *unstable set* of M (under f) (in notation, $\{x_{-n}\}_{n\in\mathbb{Z}_+} \in W^-(M)$) if $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \subset M$.

(iii) A compact positive orbit $\{x_n\}_{n\in\mathbb{Z}_+}$ is said to be in the *weakly stable set* of M (under f) (in notation, $\{x_n\}_{n\in\mathbb{Z}_+} \in W_w^+(M)$) if $\Lambda^+(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \cap M \neq \emptyset$.

(iv) A compact negative orbit $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ is said to be in the *weakly unstable set* of M (under f) (in notation, $\{x_{-n}\}_{n\in\mathbb{Z}_+} \in W_w^-(M)$) if $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \cap M \neq \emptyset$.

Under (A4), $f|_Y : Y \to Y$ defines a discrete semidynamical system. We will have occasion to consider the dynamics of $f|_Y$. To distinguish the notions of stable sets, etc., in this case with those defined previously for f, we will employ the notation $W^+(M; f|_Y)$, etc. when we refer to $f|_Y$.

DEFINITION 2.10. Let M_1 and M_2 be two isolated invariant sets. We say that M_1 is *chained to* M_2 if there exists an orbit $\{x_n\}_{n\in\mathbb{Z}}$ with $x_k \notin M_1 \cup M_2$ for some $k \in \mathbb{Z}$ such that $\{x_{-n}\}_{n\in\mathbb{Z}_+} \in W^-(M_1)$ and $\{x_n\}_{n\in\mathbb{Z}_+} \in W^+(M_2)$.

DEFINITION 2.11. A finite sequence $M_1, M_2, \cdots, M_k$ of isolated invariant sets will be called a *chain* if $M_1 \to M_2 \to \cdots \to M_k$. A chain is called a *cycle* if $M_k = M_1$.

In the following definitions, we will assume that (A3)-(A5) hold.

DEFINITION 2.12. A covering $\Pi = \{M_1, M_2, \cdots, M_k\}$ of $\bar{\Omega}(f|_Y)$ is called an *isolated covering* of $f|_Y$ if $M_1, M_2, \cdots, M_k$ are pairwise disjoint, compact and isolated invariant.

Note that each $M_i \subset Y$ $(i = 1, \cdots, k)$ is required to be an isolated invariant set in X under f and that $\bar{\Omega}(f|_Y) \subset \cup_{i=1}^k M_i$.

DEFINITION 2.13. An isolated covering $\Pi = \{M_1, \cdots, M_k\}$ of $f|_Y$ is called an *acyclic covering* of $f|_Y$ if no subsets of Π form a cycle for $f|_Y$ in Y.

Note that the "acyclic" condition is a requirement on an isolated covering of $f|_Y$ but not on $f|_Y$ itself.

In the statements of our main results on persistence in the next section, we always assume (A1)-(A6) and the following hypotheses:

(H1) $f|_Y$ is dissipative.

(H2) $f|_Y$ has an acyclic covering $\Pi = \{M_1, M_2, \cdots, M_k\}$.

PROPOSITION 2.2. *Let* (A1)-(A6) *and* (H1), (H2) *hold. Then for any positive orbit* $\{x_n\}_{n\in\mathbb{Z}_+}$ *in* Y, *there exists one and only one* i *such that* $\{x_n\}_{n\in\mathbb{Z}_+} \in W^+(M_i; f|_Y)$.

Proof. Since the positive orbit $\{x_n\}_{n\in\mathbb{Z}_+}$ is compact, by Proposition 2.1, $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+})$ is compactly invariantly connected. On the other hand, $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) \subset \Omega(f|_Y) \subset \cup_{i=1}^k M_i$. Thus, $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) \subset M_i$ for some i and $\{x_n\}_{n\in\mathbb{Z}_+} \in W^+(M_i)$. □

3. Persistence. Various forms of the following theorem have been referred to as the Butler-McGehee lemma. In the original form (Freedman and Waltman [7, App. A, Lemma 1]), the setting is for a hyperbolic restpoint of an autonomous ordinary differential equation. Since then it has been extended to an isolated invariant set for a continuous flow on a locally compact metric space by Butler and Waltman [2] and for continuous semiflows by Dunbar, Rybakowski, and Schmitt [3]. Our version here is similar to that in [2] except that now we have a discrete semiflow.

THEOREM 3.1. *Let X be a metric space with metric d. Let $f: X \to X$ be continuous and let M be an isolated invariant set in X.*

(I) *If $\{x_n\}_{n\in\mathbb{Z}^+}$ is a compact positive orbit and $\{x_n\}_{n\in\mathbb{Z}^+} \in W_w^+(M)\setminus W^+(M)$, then*
  (a) *There exists a positive orbit $\{y_n\}_{n\in\mathbb{Z}^+}$ in $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$ such that $y_0 \notin M$ and $\{y_n\}_{n\in\mathbb{Z}^+} \in W^+(M)$, and*
  (b) *There exists a negative orbit $\{z_{-n}\}_{n\in\mathbb{Z}^+}$ in $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$ such that $z_0 \notin M$ and $\{z_{-n}\}_{n\in\mathbb{Z}^+} \in W^-(M)$.*

(II) *If $\{x_{-n}\}_{n\in\mathbb{Z}^+}$ is a compact negative orbit and $\{x_{-n}\}_{n\in\mathbb{Z}^+} \in W_w^-(M)\setminus W^-(M)$, then*
  (a) *There exists a positive orbit $\{y_n\}_{n\in\mathbb{Z}^+}$ in $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}^+})$ such that $y_0 \notin M$ and $\{y_n\}_{n\in\mathbb{Z}^+} \in W^+(M)$, and*
  (b) *There exists a negative orbit $\{z_{-n}\}_{n\in\mathbb{Z}^+}$ in $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}^+})$ such that $z_0 \notin M$ and $\{z_{-n}\}_{n\in\mathbb{Z}^+} \in W^-(M)$.*

Proof. (I)(a). It suffices to show that there exists a positive orbit $\{y_n\}_{n\in\mathbb{Z}^+}$ in $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$ such that $y_0 \notin M$ and $y_n \in N$ for all $n \in \mathbb{Z}_+$ sufficiently large, where N is any closed isolating neighbourhood of M. (Since then $\Lambda^+(\{y_n\}_{n\in\mathbb{Z}^+}) \subset N$ and $\Lambda^+(\{y_n\}_{n\in\mathbb{Z}^+})$ is invariant, this implies $\Lambda^+(\{y_n\}_{n\in\mathbb{Z}^+}) \subset M$ and hence $\{y_n\}_{n\in\mathbb{Z}^+} \in W^+(M)$.) Since $\{x_n\}_{n\in\mathbb{Z}^+} \in W_w^+(M)\setminus W^+(M)$, there exist $p_1, p_2 \in \Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$ such that $p_1 \in M$ and $p_2 \notin M$. Let N be a closed isolating neighbourhood of M such that $p_2 \notin N$. Let $\{x_{1_n}\}_{n\in\mathbb{Z}^+}$ and $\{x_{2_n}\}_{n\in\mathbb{Z}^+}$ be two (convergent) subsequences of $\{x_n\}_{n\in\mathbb{Z}^+}$ such that $\lim_{n\to\infty} x_{1_n} = p_1$ and $\lim_{n\to\infty} x_{2_n} = p_2$. By going to subsequences if necessary, we can assume that $x_{1_n} \in \operatorname{int}(N)$, $x_{2_n} \notin N$, and $2_n < 1_n < 2_{n+1}$ for all $n \in \mathbb{Z}_+$. Define a subsequence $\{x_{3_n}\}_{n\in\mathbb{Z}^+}$ of $\{x_n\}_{n\in\mathbb{Z}^+}$ by the properties: (i) $2_n < 3_n \leqq 1_n$, (ii) $x_{3_n-1} \notin N$, and (iii) $x_k \in N$ for all k, $3_n \leqq k \leqq 1_n$. Again by going to subsequences if necessary, we can assume that $\lim_{n\to\infty} x_{3_n-1} = y_0$ exists. Then $\lim_{n\to\infty} x_{3_n} = y_1$ also exists. Moreover, $y_0 \notin \operatorname{int}(N)$ (and hence $\notin M$), $y_1 \in N$, and $y_1 = f(y_0)$. Let $\{y_n\}_{n\in\mathbb{Z}^+}$ be the positive orbit through y_0. Then $\{y_n\}_{n\in\mathbb{Z}^+} \subset \Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$. There are two cases to consider.

Case 1. $\{1_n - 3_n\}_{n\in\mathbb{Z}_+}$ is unbounded. By going to subsequences if necessary, we can assume that $\{1_n - 3_n\}_{n\in\mathbb{Z}_+} \uparrow \infty$. Fix any $k \in \mathbb{Z}_+$, $k \geqq 1$. Since $y_k = \lim_{n\to\infty} x_{3_n+k-1}$ and $3_n \leqq 3_n + k - 1 \leqq 1_n$ for $n \in \mathbb{Z}_+$ sufficiently large, we have $y_k \in N$. Thus, $\Lambda^+(\{y_n\}_{n\in\mathbb{Z}^+}) \subset N$ and $\{y_n\}_{n\in\mathbb{Z}^+} \in W^+(M)$.

Case 2. $\{1_n - 3_n\}_{n\in\mathbb{Z}_+}$ is bounded. By going to subsequences if necessary, we can assume that $1_n - 3_n = m$ for some $m \in \mathbb{Z}_+$. Therefore, $y_{m+1} = \lim_{n\to\infty} x_{3_n+m+1-1} = \lim_{n\to\infty} x_{1_n} = p_1 \in M$. Since M is invariant, $y_n \in M$ for $n \in \mathbb{Z}_+$, $n > m$. Thus, again $\{y_n\}_{n\in\mathbb{Z}^+} \in W^+(M)$.

(I)(b). It suffices to show that there exists a negative orbit $\{z_{-n}\}_{n\in\mathbb{Z}^+}$ in $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$ such that z_0 is not in M and $z_{-n} \in N$ for all $n \in Z_+$ sufficiently large, where N is any closed isolating neighbourhood of M. (Since then $\Lambda^-(\{z_{-n}\}_{n\in\mathbb{Z}^+}) \subset N$ and $\Lambda^-(\{z_{-n}\}_{n\in\mathbb{Z}^+})$ is invariant, this implies $\Lambda^-(\{z_{-n}\}_{n\in\mathbb{Z}^+}) \subset M$ and hence $\{z_{-n}\}_{n\in\mathbb{Z}^+} \in W^-(M)$.) Since $\{x_n\}_{n\in\mathbb{Z}^+} \in W_w^+(M)\setminus W^+(M)$, there exist $p_1, p_2 \in \Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$ such that $p_1 \in M$ and $p_2 \notin M$. Let N be a closed isolating neighbourhood of M such that $p_2 \notin N$. Let $\{x_{1_n}\}_{n\in\mathbb{Z}^+}$ and $\{x_{2_n}\}_{n\in\mathbb{Z}^+}$ be two subsequences of $\{x_n\}_{n\in\mathbb{Z}^+}$ such that $\lim_{n\to\infty} x_{1_n} = p_1$ and $\lim_{n\to\infty} x_{2_n} = p_2$. By going to subsequences if necessary, we can assume that $x_{1_n} \in \operatorname{int}(N)$, $x_{2_n} \notin N$, and $1_n < 2_n < 1_{n+1}$ for all $n \in \mathbb{Z}_+$. Define a subsequence $\{x_{3_n}\}_{n\in\mathbb{Z}^+}$ of $\{x_n\}_{n\in\mathbb{Z}^+}$

by the following properties: (i) $1_n \leqq 3_n < 2_n$; (ii) $x_{3_n+1} \notin N$; and (iii) $x_k \in N$ for all k, $1_n \leqq k \leqq 3_n$. Again by going to subsequences if necessary, we can assume that $\lim_{n\to\infty} x_{3_n} = z_{-1}$ exists. Then $\lim_{n\to\infty} x_{3_n+1} = z_0$ also exists. Moreover, $z_0 \notin \text{int}(N)$ (and hence $\notin M$), $z_{-1} \in N$, and $z_0 = f(z_{-1})$. We construct a negative orbit $\{z_{-n}\}_{n\in\mathbb{Z}^+}$ through z_0 in $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}^+})$ as follows. z_{-2} is the limit of any convergent subsequence $\{x_{(-2)_n}\}_{n\in\mathbb{Z}^+}$ of $\{x_{3_n-1}\}_{n\in\mathbb{Z}^+}$, z_{-3} is the limit of any convergent subsequence $\{x_{(-3)_n}\}_{n\in\mathbb{Z}^+}$ of $\{x_{(-2)_n-1}\}_{n\in\mathbb{Z}^+}$, etc. There are two cases to consider.

Case 1. $\{3_n - 1_n\}_{n\in\mathbb{Z}_+}$ is unbounded. By going to subsequences if necessary, we can assume that $\{3_n - 1_n\}_{n\in\mathbb{Z}_+} \uparrow \infty$. We will show that $z_{-k} \in N$ for all $k \in \mathbb{Z}_+$, $k \geqq 1$. Fix any $k \in \mathbb{Z}_+$, $k \geqq 1$. Recall that $z_{-k} = \lim_{n\to\infty} x_{(-k)_n}$. Since $\{(-k)_n\}_{n\in\mathbb{Z}_+}$ is a subsequence of $\{3_n - k + 1\}_{n\in\mathbb{Z}_+}$ and $1_n \leqq 3_n - k + 1 \leqq 3_n$ for $n \in \mathbb{Z}_+$ sufficiently large, therefore $x_{(-k)_n} \in N$ for $n \in \mathbb{Z}_+$ sufficiently large. Thus $z_{-k} \in N$. Hence, $\Lambda^-(\{z_{-n}\}_{n\in\mathbb{Z}^+}) \subset N$ and $\{z_{-n}\}_{n\in\mathbb{Z}^+} \in W^-(M)$.

Case 2. $\{1_n - 3_n\}_{n\in\mathbb{Z}_+}$ is bounded. By going to subsequences if necessary, we can assume that $3_n - 1_n = m$ for some $m \in \mathbb{Z}_+$. Now, $z_0 = \lim_{n\to\infty} x_{3_n+1} = \lim_{n\to\infty} x_{1_n+m+1} = f^{m+1}(p_1) \in M$, since $p_1 \in M$ and M is invariant, which is a contradiction. This shows that $\{1_n - 3_n\}_{n\in\mathbb{Z}_+}$ must be unbounded.

The proofs of (II)(a) and (II)(b) are similar to (I)(a) and (I)(b). □

The following proposition together with Proposition 2.2 shows that an acyclic covering of $f|_Y$ is a Morse decomposition for $\Omega(f|_Y)$.

PROPOSITION 3.2. *Let* (A1)-(A6) *and* (H1), (H2) *hold. Then for any compact negative orbit* $\{x_{-n}\}_{n\in\mathbb{Z}_+}$ *in* Y, *there exists one and only one* i *such that* $\{x_{-n}\}_{n\in\mathbb{Z}_+} \in W^-(M_i; f|_Y)$.

Proof. Note that we are only interested in $f|_Y$ and not f here. Suppose $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \not\subset M_i$ for all i. Let $y \in \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$. Since $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ is invariant, by (A4) and (A5), there exists an orbit $\{y_n\}_{n\in\mathbb{Z}}$ through y such that $\{y_n\}_{n\in\mathbb{Z}}$ is in Y and $\{y_n\}_{n\in\mathbb{Z}} \subset \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$. Thus, $\emptyset \neq \Lambda^+(\{y_n\}_{n\in\mathbb{Z}_+}) \subset \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$. Moreover, $\Lambda^+(\{y_n\}_{n\in\mathbb{Z}_+}) \subset M_{i_1}$ for some i_1, by Proposition 2.2. Thus, $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \cap M_{i_1} \neq \emptyset$, that is, $\{x_{-n}\}_{n\in\mathbb{Z}_+} \in W_w^-(M_{i_1}; f|_Y)$. Since $\{x_{-n}\}_{n\in\mathbb{Z}_+} \notin W^-(M_{i_1}; f|_Y)$ by assumption, by Theorem 3.1 there exists a negative orbit $\{z_{-n}\}_{n\in\mathbb{Z}^+} \subset \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ such that $z_0 \notin M_{i_1}$ and $\{z_{-n}\}_{n\in\mathbb{Z}^+} \in W^-(M_{i_1}; f|_Y)$. Let $\{z_n\}_{n\in\mathbb{Z}}$ be the (full) orbit extending $\{z_{-n}\}_{n\in\mathbb{Z}^+}$. By Proposition 2.2 again, there exists i_2 such that $\{z_n\}_{n\in\mathbb{Z}^+} \in W^+(M_{i_2}; f|_Y)$. Clearly, there exists n such that $z_n \notin M_{i_2}$. Therefore $M_{i_1} \to M_{i_2}$. Since $\emptyset \neq \Lambda^+(\{z_n\}_{n\in\mathbb{Z}_+}) \subset \Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+})$ and $\Lambda^+(\{z_n\}_{n\in\mathbb{Z}_+}) \subset M_{i_2}$, $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \cap M_{i_2} \neq \emptyset$, that is, $\Lambda^-(\{x_{-n}\}_{n\in\mathbb{Z}_+}) \in W_w^-(M_{i_2}; f|_Y)$. Since $\{x_{-n}\}_{n\in\mathbb{Z}_+} \notin W^-(M_{i_2}; f|_Y)$, we can repeat the above argument to get an i_3 such that $M_{i_2} \to M_{i_3}$. Since there are only a finite number of M_i's, we will eventually arrive at a cycle. This contradicts (H2). □

THEOREM 3.3. *Under assumptions* (A1)-(A6) *and hypotheses* (H1) *and* (H2), f *is persistent* (*with respect to* Y) *if and only if we have the following*:

(H3) *There is no positive orbit* $\{x_n\}_{n\in\mathbb{Z}_+}$ *in* $X \setminus Y$ *such that* $\{x_n\}_{n\in\mathbb{Z}_+} \in W^+(M_i)$ *for some* i (*in notation, we write* $W^+(M_i) \cap X \setminus Y = \emptyset$ *for all* $i = 1, 2, \cdots, k$).

Proof. (→ part.) If (H3) fails to hold, then there exists a positive orbit $\{x_n\}_{n\in\mathbb{Z}_+}$ in $X \setminus Y$ such that $\{x_n\}_{n\in\mathbb{Z}_+} \in W^+(M_i)$ for some i. Then $\lim_{n\to\infty} d(x_n, M_i) = 0$ and $\limsup_{n\to\infty} d(f^n(x), Y) = 0$, where $x = x_0$. Thus, f is not weakly persistent (with respect to Y) and hence not persistent (with respect to Y).

(← part.) Suppose f is not persistent (with respect to Y). Then there exists a positive orbit $\{x_n\}_{n\in\mathbb{Z}_+} \subset X \setminus Y$ such that $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) \cap Y \neq \emptyset$. Let $y \in \Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) \cap Y$ and denote the positive orbit through y by $\{y_n\}_{n\in\mathbb{Z}_+}$. Then $\{y_n\}_{n\in\mathbb{Z}_+} \subset Y$ and $\emptyset \neq \Lambda^+(\{y_n\}_{n\in\mathbb{Z}_+}) \subset \Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) \cap Y$. By Proposition 3.2, there exists i_1 such that

$\Lambda^+(\{y_n\}_{n\in\mathbb{Z}_+}) \subset M_{i_1}$. Thus, $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+}) \cap M_{i_1} \neq \emptyset$, that is, $\{x_n\}_{n\in\mathbb{Z}_+} \in W_w^+(M_{i_1})$. Since, by (H3), $\{x_n\}_{n\in\mathbb{Z}_+} \notin W^+(M_{i_1})$, we can apply Theorem 3.1 to get a positive orbit $\{z_n\}_{n\in\mathbb{Z}_+} \subset \Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+})$ such that $z_0 \notin M_{i_1}$ and $\{z_n\}_{n\in\mathbb{Z}_+} \in W^+(M_{i_1})$. By (H3), $\{z_n\}_{n\in\mathbb{Z}_+} \subset Y$. Clearly $z_0 \notin M_i$ for all i. Let $\{z_n\}_{n\in\mathbb{Z}}$ be an orbit through z_0. It exists because $z_0 \in \Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+})$ and $\Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+})$ is invariant. Clearly, $\{z_n\}_{n\in\mathbb{Z}} \subset Y$ and $\{z_n\}_{n\in\mathbb{Z}} \subset \Lambda^+(\{x_n\}_{n\in\mathbb{Z}_+})$. Since $\{z_{-n}\}_{n\in\mathbb{Z}_+}$ is a compact negative orbit, by Proposition 3.2, $\{z_{-n}\}_{n\in\mathbb{Z}_+} \in W^-(M_{i_2})$ for some i_2. Therefore, $M_{i_2} \to M_{i_1}$. Clearly, $\{x_n\}_{n\in\mathbb{Z}_+} \in W_w^+(M_{i_2})$ and we can repeat the above argument to get an i_3 such that $M_{i_3} \to M_{i_2}$. Since there are only a finite number of M_i's, we will eventually arrive at a cycle. This contradicts (H2). □

4. Application to a predator-prey model. It has been shown that models of arthropod parisitoid-host systems [8] and models of predator-prey systems when population numbers are small [5] can be best described by a discrete semidynamical system. In this section we describe such a system, which includes models considered in [9]-[11] as special cases.

Let x_n be the prey or host population numbers in the nth generation and let y_n be the predator or parasitoid population. Then the model takes the form

$$\begin{aligned} x_{n+1} &= x_n F(x_n, y_n), & x_0 &\geqq 0, \\ y_{n+1} &= x_n y_n G(x_n, y_n), & y_0 &\geqq 0. \end{aligned} \tag{4.1}$$

In this case, $X = \mathbb{R}_+^2 = \{(x, y): x \geqq 0, y \geqq 0\}$ and $Y = \{(x, 0): x \geqq 0\} \cup \{(0, y): y \geqq 0\}$, the nonnegative x and y axes. Moreover, persistence is synonymous with the survival of both populations over all time.

The mathematical assumptions and their biological interpretations, where appropriate, are as follows:

(PP1) $F, G \in \mathscr{C}([0, \infty) \times [0, \infty))$. Small changes in population numbers in any generation result in small changes in the succeeding generation.

(PP2) $F(0, 0) > 1$. In the absence of predators, the prey population will grow if its numbers are small.

(PP3) $F(x, y) \geqq 0$, $G(x, y) \geqq 0$ if $x \geqq 0$, $y \geqq 0$; $F(x, y) > 0$, $G(x, y) > 0$ if $x > 0$, $y > 0$. In general there cannot be a negative number of populations. Further, if both populations are present, extinction of either population cannot occur in finite time.

(PP4) $F(x, y)$ is monotonically decreasing in y. Further, $\lim_{y\to\infty} F(x, y) = 0$. The larger the number of predators, the smaller the growth rate of the prey. For a very large number of predators, the prey population is driven to near extinction.

(PP5) There exists $x_K > 0$ such that $F(x_K, 0) = 1$ and $F(x, 0) < 1$ for all $x > x_K$. There is a carrying capacity of the environment beyond which the prey population cannot be sustained, even in the absence of predators.

(PP6) $xG(x, y)$ is a strictly increasing function of x with $x_K G(x_K, 0) > 1$. The more food for the predator, the higher will be its growth rate. Further, at some prey value prior to carrying capacity, and thereafter, the predator population can sustain itself on the prey population.

(PP7) $G(x, y)$ is monotonically decreasing in y. The per capita growth rate of the predator is diminished for larger values of its population due to intraspecific competition for its food.

Before stating the final hypothesis for model (4.1), we need to study the single-species dynamics of the prey in the absence of predators. This dynamics is given by the equation

$$x_{n+1} = x_n F(x_n, 0), \qquad x_0 \geqq 0. \tag{4.2}$$

The dynamics of system (4.2) is discussed in [6]. Let $\tilde{x} = \sup_{0 \leqq x \leqq x_K} xF(x, 0)$. Then $\tilde{x} \geqq x_K$ and if $\{x_n\}_{n \in \mathbb{Z}_+}$ is a positive orbit of (4.2), then $\limsup_{n \to \infty} x_n \leqq \tilde{x}$.

Further, assuming (PP1), (PP2), (PP5), it is shown in [6] that there exists $\eta > 0$, η computable, such that for $x_0 > 0$, $\Lambda^+(\{x_n\}_{n \in \mathbb{Z}_+}) \subset [\eta, \tilde{x}]$.

The final hypothesis is the following.

(PP8) $\eta G(\eta, 0) > 1$.

By (PP6), (PP8) implies $xG(x, 0) > 1$ for all $x \geqq \eta$, which in turn implies the semi-axis $\{(x, 0) \in \mathbb{R}^2_+ : x \geqq \eta\}$ repels in the y direction.

To show that system (4.1) is persistent under assumptions (PP1)-(PP8), we need to check that (A1)-(A6) and (H1)-(H3) are satisfied.

Clearly (A1)-(A5) are satisfied. To show (A6) and (H1), let $\{(x_n, y_n)\}_{n \in \mathbb{Z}_+}$ be a positive orbit of (4.1) in $\mathbb{R}^2_+$. Then since $F(x, y) \leqq F(x, 0)$, $\limsup_{n \to \infty} x_n \leqq \tilde{x}$. For the purpose of the following discussion we can assume that $x_n \leqq \bar{x}$ for all n, where $\bar{x} \equiv \tilde{x} + 1$.

Let $\phi(y) = \sup_{0 \leqq x \leqq \bar{x}} xF(x, y)$. By (PP4), $\phi(y)$ is a decreasing function of y and $\lim_{y \to \infty} \phi(y) = 0$. Let $\delta > 0$ be such that

$$\delta \bar{x} G(\bar{x}, 0) G(\delta, 0) = 1 \tag{4.3}$$

which exists by (PP6) since $\bar{x}G(\bar{x}, 0) > 1$. Further, choose $\tilde{y}$ so large that $\phi(y) < \delta$ for all $y \geqq \tilde{y}$ and define $\hat{y} = \bar{x}\tilde{y}G(\bar{x}, 0)$. Clearly $\hat{y} > \tilde{y}$. Moreover, $x_{n+1} = x_n F(x_n, y_n) \leqq \phi(y_n)$.

Let us make the following observations.

Observation 1. Suppose y_n is such that $0 < y_n \leqq \tilde{y}$. Then $y_{n+1} = x_n y_n G(x_n, y_n) \leqq \bar{x}G(\bar{x}, 0) y_n \leqq \bar{x}G(\bar{x}, 0)\tilde{y} = \hat{y}$.

Observation 2. If y_n is such that $y_n > \tilde{y}$, then $x_{n+1} \leqq \phi(y_n) < \delta$ and $y_{n+2} = x_{n+1} y_{n+1} G(x_{n+1}, y_{n+1}) = x_{n+1} x_n y_n G(x_n, y_n) G(x_{n+1}, y_{n+1}) < \delta \bar{x} G(\bar{x}, 0) G(\delta, 0) y_n = y_n$, by (4.3).

We now consider two cases.

Case 1. $y_n > \hat{y}$ for all n. In this case, by Observation 2, $y_0 > y_2 > y_4 > \cdots$. Let $y_* = \lim_{k \to \infty} y_{2k}$. Since $y_{2k+1} \leqq \bar{x}G(\bar{x}, 0) y_{2k}$, $\limsup_{n \to \infty} y_n \leqq y_* \bar{x} G(\bar{x}, 0)$. Claim: $y_* = \hat{y}$. Suppose not, that is, suppose $y_* > \hat{y}$. Let $\{x_{2k_j}\}$ be a convergent subsequence of $\{x_{2k}\}$ and denote its limit by x_*. Also denote the map on the right-hand side of (4.1) by f. Then $f^2(x_*, y_*) = f^2(\lim_{j \to \infty} x_{2k_j}, \lim_{j \to \infty} y_{2k_j})$ so that the second component of $f^2(x_*, y_*)$ is $\lim_{j \to \infty} y_{2k_j+2} = y_*$. This contradicts Observation 2 since $y_* > \hat{y} > \tilde{y}$. Hence, we have shown that $\limsup_{n \to \infty} y_n \geqq \hat{y}\bar{x}G(\bar{x}, 0)$.

Case 2. There exists n_0 such that $y_{n_0} \leqq \hat{y}$. If $y_{n_0} \leqq \tilde{y}$, by Observation 1, $y_{n_0+1} \leqq \hat{y}$. If $\tilde{y} < y_{n_0} \leqq \hat{y}$, then $y_{n_0+1} \leqq \hat{y}\bar{x}G(\bar{x}, 0)$ and, by Observation 2, $y_{n_0+2} \leqq y_{n_0} \leqq \hat{y}$. Proceeding inductively, we can easily show $y_n \leqq \hat{y}\bar{x}G(\bar{x}, 0)$ for all $n \geqq n_0$ and thus $\limsup_{n \to \infty} y_n \leqq \hat{y}\bar{x}G(\bar{x}, 0)$.

This shows that $\Omega(f) \subset [0, \bar{x}] \times [0, \hat{y}\bar{x}G(\bar{x}, 0)]$ and hence (A6) and (H1) are satisfied.

To show (H2) is satisfied we note that $\{M_1, M_2\}$ is an isolated covering of $f|_Y$, where $M_1 = \{(0, 0)\}$ and M_2 is the maximal invariant set in $\{(x, 0) : \eta \leqq x \leqq \tilde{x}\}$, and that this covering is acyclic.

Finally, we show (H3). We first show that $W^+(M_1) \cap \operatorname{int}(\mathbb{R}^2_+) = \varnothing$. Suppose not. Let $\{(x_n, y_n)\}_{n \in \mathbb{Z}_+}$ be a positive orbit in $\operatorname{int}(\mathbb{R}^2_+)$ such that $\lim_{n \to \infty} (x_n, y_n) = (0, 0)$. Since $F(0, 0) > 1$, $F(x, y) > 1$ for (x, y) near $(0, 0)$. Thus for sufficiently large n, $x_{n+1} =$

$x_nF(x_n, y_n) > x_n$, contradicting $\lim_{n\to\infty} x_n = 0$. Next, we show that $W^+(M_2) \cap \text{int}(\mathbb{R}^2_+) = \varnothing$. Again, suppose not. Let $\{(x_n, y_n)\}_{n\in\mathbb{Z}_+}$ be a positive orbit in $\text{int}(\mathbb{R}^2_+)$ such that $\lim_{n\to\infty} d((x_n, y_n), M_2) = 0$. Then $\liminf_{n\to\infty} x_n \geqq \eta$, $\limsup_{n\to\infty} x_n \leqq \tilde{x}$, and $\lim_{n\to\infty} y_n = 0$. Since by (PP8) $\eta G(\eta, 0) > 1$, let $\varepsilon > 0$ be such that $(1-\varepsilon)\eta G(\eta, 0) > 1$. Since G is continuous, for sufficiently large n, we have $G(x, y_n) \geqq (1-\varepsilon)G(x, 0)$ for all $0 \leqq x \leqq \tilde{x}+1$. Thus, $y_{n+1} = x_n y_n G(x_n, y_n) \geqq (1-\varepsilon)x_n G(x_n, 0)y_n > y_n$ for sufficiently large n, contradicting $\lim_{n\to\infty} y_n = 0$.

Hence all the hypotheses for Theorem 3.3 are satisfied and model (4.1) exhibits persistence.

5. Discussion. The notion of persistence, originating in the theory of dynamical systems in locally compact metric spaces [1], [2], has been extended in various manners in recent work. In this paper we have extended the notion of persistence to discrete semidynamical systems. To obtain testable persistence criteria, we have proved a Butler-McGehee type lemma for such systems.

In [6], we have obtained testable uniform persistence criteria for one-dimensional maps. Here, we extend that work to obtain testable criteria for persistence in higher-dimensional systems.

We have applied our results to a class of discrete predator-prey models, which have been discussed in the literature as to their relevance, but not as to persistence and extinction of the modeled populations. In particular, a feature of these models is that all solutions initiating on the y-axis map to the origin in one iteration (in the absence of food, all predators die). By continuity, solutions initiating close to the y-axis are mapped to a neighbourhood of the origin. Hence, it is not obvious that the omega limit set of such an orbit is bounded away from the axes (persistence). Here we give criteria, with reasonable biological interpretations, for persistence to occur.

REFERENCES

[1] G. J. BUTLER, H. I. FREEDMAN, AND P. WALTMAN, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425-430.

[2] G. J. BUTLER AND P. WALTMAN, *Persistence in dymamical systems*, J. Differential Equations, 63 (1986), pp. 255-263.

[3] S. R. DUNBAR, K. P. RYBAKOWSKI, AND K. SCHMITT, *Persistence in models of predator-prey populations with diffusion*, J. Differential Equations, 65 (1986), pp. 117-138.

[4] A. FONDA, *Uniformly persistent semi-dynamical systems*, Proc. Amer. Math. Soc., 104 (1988), pp. 111-116.

[5] H. I. FREEDMAN, *Deterministic Mathematical Models in Population Ecology*, Marcel Dekker, New York, 1980.

[6] H. I. FREEDMAN AND J. W.-H. SO, *Persistence in discrete models of a population which may be subjected to harvesting*, Natur. Resource Modeling, 2 (1987), pp. 135-145.

[7] H. I. FREEDMAN AND P. WALTMAN, *Persistence in models of three interacting predator-prey populations*, Math. Biosci., 68 (1984), pp. 213-231.

[8] M. P. HASSELL, *The dynamics of arthropod predator-prey systems*, Princeton University Press, Princeton, NJ, 1978.

[9] M. P. HASSELL AND H. N. COMINS, *Discrete time models for two-species competition*, Theoret. Population Biol., 9 (1976), pp. 202-221.

[10] M. P. HASSELL AND R. M. MAY, *Stability in insect host-parasite models*, J. Animal Ecology, 42 (1973), pp. 693-726.

[11] M. P. HASSELL, J. K. WAAGE, AND R. M. MAY, *Variable parasitoid sex ratios and their effect on host-parasitoid dynamics*, J. Animal Ecology, 52 (1983), pp. 889-904.

[12] J. P. LASALLE, *The Stability of Dynamical Systems*, Society for Industrial and Applied Mathematics, Philadelphia, 1976.

[13] V. V. NEMYITSKII AND V. V. STEPANOV, *Qualitative theory of differential equations*, Princeton University Press, Princeton, NJ, 1960.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 939-949, July 1989

010

OSCILLATION AND NONOSCILLATION FOR SYSTEMS OF SELF-ADJOINT SECOND-ORDER DIFFERENCE EQUATIONS*

SHAOZHU CHEN† AND LYNN H. ERBE‡

Abstract. The self-adjoint second-order difference system (1) $\Delta(R_n\Delta Y_n)+P_nY_{n+1}=0$, $n\geqq 0$, where $\Delta Y_n = Y_{n+1}-Y_n$, $\{R_n\}_{n=0}^{\infty}$, $\{P_n\}_{n=0}^{\infty}$ are sequences of $d\times d$ Hermitian matrices with $R_n>0$ (positive definite) are considered. Oscillation and nonoscillation criteria for solutions of (1) are obtained by Riccati and averaging techniques.

Key words. difference equation, oscillation, nonoscillation, Riccati, averaging

AMS(MOS) subject classification. 39A10

1. Introduction.

Consider the self-adjoint second-order difference system

$$\Delta(R_n\Delta Y_n)+P_nY_{n+1}=0, \qquad n\geqq 0, \tag{1.1}$$

where Δ is the forward difference operator $\Delta Y_n = Y_{n+1}-Y_n$, and $R=\{R_n\}_{n=0}^{\infty}$, $P=\{P_n\}_{n=0}^{\infty}$ are sequences of $d\times d$ Hermitian matrices with $R_n>0$. We remark that Hermitian matrix inequalities $A>0$ ($\geqq 0$) are in the sense of positive (nonnegative) definiteness. In this paper we are interested in establishing oscillation and nonoscillation properties of solutions of (1.1) that may be considered as discrete versions of results for the second-order linear differential system

$$(R(t)Y'(t))'+P(t)Y(t)=0, \tag{1.2}$$

where $R(t)$, $P(t)$ are continuous Hermitian-matrix valued functions defined on $[0,+\infty)$ with $R(t)>0$. This problem has attracted substantial interest recently (we refer to [1]-[4], [8] for a careful discussion of disconjugacy and related properties of (1.1) and its relation to the continuous case (1.2)). We refer to [5]-[7] and [9] for additional results in the scalar case. In particular, in [5] the authors used Riccati and averaging techniques for the scalar case of (1.1) (R_n, P_n real numbers) to obtain a number of oscillation results. These, in turn, may be considered analogues (and, in some cases, a strengthening) of criteria in the scalar case of (1.2).

The results involve "averaging" methods that have previously been used in scalar differential and difference equations which allow us to obtain more sensitive tests for oscillation. As a corollary we obtain the discrete analogue of a recent test for oscillation of matrix differential equations obtained in [4] and [8] (cf. Corollary 2.16) which has generated considerable activity. To be specific, if $R_n=I$ (the $d\times d$ identity matrix) for all n, and if $\lim_{n\to\infty}\sum_{j=0}^{n}P_j$ does not exist (finite), then

$$\Delta^2 Y_n+P_nY_{n+1}=0$$

is oscillatory. The same conclusion holds if

$$\limsup_{n\to\infty}\lambda_1\left(\sum_{j=0}^{n}P_j\right)=+\infty,$$

* Received by the editors February 24, 1988; accepted for publication (in revised form) October 26, 1988.

† Shandong University, Jinan, Shandong, People's Republic of China. Present address, Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

‡ Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The work of this author was supported by the Natural Science and Engineering Research Council of Canada.

where $\lambda_1(\cdot)$ denotes the largest eigenvalue. (This is the extension of the result of [4] and [8] referred to above.)

Let $U=\{U_n\}$, $V=\{V_n\}$ be two sequences of $d\times d$ matrices. We define the "bracket function" by

$$\{U, V\}_n = U_n^* R_n \Delta V_n - (R_n \Delta U_n)^* V_n, \qquad n \geqq 0, \tag{1.3}$$

where * denotes the conjugate transpose. It is straightforward to establish that

$$\{U, V\}_n = -\{V, U\}_n^* \tag{1.4}$$

and if U, V are solutions of (1.1), then

$$\Delta\{U, V\}_n = 0, \tag{1.5}$$

so that $\{U, V\}_n \equiv C$ (a constant matrix).

We say that a solution Y of (1.1) is *prepared* in case $\{Y, Y\}_n = 0$, $n \geqq 0$. We refer to [1] and [2] for an additional discussion of this and remark only that the analogue of the "preparedness" for (1.2) is $Y^*(t)R(t)Y'(t) \equiv Y^{*\prime}(t)R(t)Y(t)$, $t \geqq 0$. It is easy to establish that a solution Y of (1.1) is prepared if and only if $Y_{n+1}^* R_n Y_n$ is Hermitian for each $n \geqq 0$. We also have the following result whose proof is omitted.

LEMMA 1.1. *Let Y be a prepared solution of* (1.1) *and assume* $\det Y_n \neq 0$ *for $n \geqq N$ for some $N \geqq 0$. Then $R_n Y_{n+1} Y_n^{-1}$ is Hermitian for each $n \geqq N$.*

We say that a prepared solution Y of (1.1) is *nonoscillatory* if there exists an $N \geqq 0$ such that

$$Y_{n+1}^* R_n Y_n > 0 \quad \text{for } n \geqq N; \tag{1.6}$$

and Y is said to be *oscillatory* otherwise.

The next result gives a simple characterization of nonoscillatory solutions of (1.1).

LEMMA 1.2. *Suppose that Y is a prepared solution of* (1.1). *Then Y is nonoscillatory if and only if there exists an $N \geqq 0$ such that*

$$R_n Y_{n+1} Y_n^{-1} > 0 \quad \textit{for } n \geqq N. \tag{1.7}$$

Proof. Suppose that Y is nonoscillatory. Then $Y_n^* R_n Y_{n+1} = Y_{n+1}^* R_n Y_n$ for $n \geqq N$ and hence

$$R_n Y_{n+1} Y_n^{-1} = (Y_n^{-1})^* (Y_n^* R_n Y_{n+1}) Y_n^{-1} > 0 \quad \text{for } n \geqq N.$$

Conversely, if $R_n Y_{n+1} Y_n^{-1} > 0$ for $n \geqq N$, then

$$Y_{n+1}^* R_n Y_n = Y_n^* R_n Y_{n+1} = Y_n^* (R_n Y_{n+1} Y_n^{-1}) Y_n > 0 \quad \text{for } n \geqq N.$$

Hence Y is nonoscillatory.

Remark. It is easy to see that (1.7) is also equivalent to $R_n Y_n Y_{n+1}^{-1} > 0$ for $n \geqq N$. We also remark that by a Sturm-type separation theorem for (1.1), it follows that either all nontrivial prepared solutions of (1.1) are nonoscillatory or none of them are. To be specific, if in our notation, $D_n \equiv R_n + R_{n+1} - P_n > 0$ holds for all large n, the separation theorem is given in [1]. Moreover, if there exists a sequence $n_k \to \infty$ such that D_{n_k} is not positive definite, then it follows from Lemma 1.2 and some manipulation that there does not exist a prepared nonoscillatory solution of (1.1). We also remark that the fact that nonoscillation implies that D_n is eventually positive definite follows immediately from Corollary 3.1 of [2]. Furthermore, to avoid complications in applying the Sturm theory, it is necessary to restrict attention in this paper to solutions of (1.1) for which the $2d \times d$ partitioned matrix

$$\begin{pmatrix} Y_{n+1} \\ R_n \Delta Y_n \end{pmatrix}$$

has full rank (for some and hence all values of n).

Suppose that (1.1) is nonoscillatory and let Y be a solution with $R_n Y_{n+1} Y_n^{-1} > 0$ for $n \geqq N$ (Lemma 1.2). We define the Riccati difference operator by

$$Z_n = (R_n \Delta Y_n) Y_n^{-1}, \qquad n \geqq N. \tag{1.8}$$

Then since $Z_n = R_n Y_{n+1} Y_n^{-1} - R_n$, it follows by Lemmas 1.1 and 1.2 that Z_n is Hermitian and $Z_n > -R_n$ for $n \geqq N$. A simple computation gives

$$\Delta Z_n + Z_n (Z_n + R_n)^{-1} Z_n + P_n = 0, \qquad n \geqq N. \tag{1.9}$$

Equation (1.9) is called the *Riccati difference equation* associated with (1.1).

2. Oscillation and nonoscillation results. We shall assume henceforth in this paper that $D_n > 0$ holds for all large n. For any Hermitian matrix A, we will assume its eigenvalues $\lambda_k(A)$, $1 \leqq k \leqq d$, are ordered so that

$$\lambda_1(A) \geqq \cdots \geqq \lambda_d(A)$$

and as usual

$$\operatorname{tr} A = \sum_{i=1}^{d} \lambda_i(A), \qquad \|A\| = (\lambda_1(A^* A))^{1/2}.$$

THEOREM 2.1. *Equation* (1.1) *is nonoscillatory if and only if there exists a sequence* Z *of* $d \times d$ *Hermitian matrices with* $Z_n > -R_n$ *for* $n \geqq N$ *for some* $N \geqq 0$ *satisfying* (1.9).

Proof. We need only to prove the sufficiency. Let Z be the solution of (1.9) with $Z_n > -R_n$ for $n \geqq N$. Define Y by $Y_N = I$, the $d \times d$ identity matrix, and for $n \geqq N+1$

$$Y_n = \prod_{j=N}^{n-1} (R_j^{-1} Z_j + I) = (R_{n-1}^{-1} Z_{n-1} + I) \cdots (R_N^{-1} Z_N + I). \tag{2.1}$$

Then Y is a solution of (1.1) for $n \geqq N$ and moreover the matrix

$$Y_{n+1}^* R_n Y_n = Y_n^* (Z_n R_n^{-1} + I) R_n Y_n = Y_n^* (Z_n + R_n) Y_n$$

is Hermitian for all $n \geqq N$ and hence Y is a prepared solution of (1.1). Since $R_n Y_{n+1} Y_n^{-1} = R_n (R_n^{-1} Z_n + I) = Z_n + R_n > 0$ for $n \geqq N$, it follows by Lemma 1.2 that (1.1) is nonoscillatory. This completes the proof.

Next suppose that (1.1) is nonoscillatory and let Z be a Hermitian solution of (1.9). Then by Lemma 1.2 we have

$$\begin{aligned} Z_N - Z_N (Z_N + R_N)^{-1} Z_N &= Z_N (Z_N + R_N)^{-1} R_N = Z_N Y_N Y_{N+1}^{-1} \\ &= R_N - R_N Y_N Y_{N+1}^{-1} < R_N. \end{aligned}$$

Hence, from (1.9) for all $n \geqq N+1$ we have

$$\begin{aligned} \sum_{j=N}^{n-1} P_j &= Z_N - \sum_{j=N}^{n-1} Z_j (Z_j + R_j)^{-1} Z_j - Z_n \\ &< Z_N - Z_N (Z_N + R_N)^{-1} Z_N + R_n \\ &< R_N + R_n. \end{aligned} \tag{2.2}$$

In particular, letting $n = N+1$ and replacing N by n, we have

$$P_n < R_n + R_{n+1} \quad \text{for } n \geqq N, \tag{2.3}$$

which is the condition $D_n > 0$ (see the remark following Lemma 1.2). Actually, that (2.2), a generalization of (2.3), holds for all large N and $n \geqq N+1$ is a necessary condition for (1.1) to be nonoscillatory. This establishes Theorem 2.2.

THEOREM 2.2. *If there exist two sequences of integers* $\{N_k\}$ *and* $\{n_k\}$ *with* $n_k \geqq N_k + 1$, $N_k \to \infty$ *as* $k \to \infty$ *such that*

$$\lambda_d \left(R_{n_k} + R_{N_k} - \sum_{j=N_k}^{n_k-1} P_j \right) \leqq 0, \qquad k = 1, 2, \cdots, \tag{2.4}$$

then (1.1) *is oscillatory.*

From this we may easily obtain Corollary 2.3.

COROLLARY 2.3. *If there exist three sequences of integers* $\{N_k\}$, $\{M_k\}$, $\{n_k\}$ *with* $n_k \geqq M_k + 1$, $M_k \geqq N_k + 1$, $N_k \to \infty$ *as* $k \to \infty$ *such that*

$$\sum_{j=N_k}^{M_k-1} P_j \geqq R_{N_k} \tag{2.5}$$

and

$$\lambda_d \left(R_{n_k} - \sum_{j=M_k}^{n_k-1} P_j \right) \leqq 0, \qquad k = 1, 2, \cdots, \tag{2.6}$$

then (1.1) *is oscillatory.*

Proof. We have

$$\begin{aligned} \lambda_d \left(R_{n_k} + R_{N_k} - \sum_{j=N_k}^{n_k-1} P_j \right) &= \lambda_d \left\{ R_{n_k} - \sum_{j=M_k}^{n_k-1} P_j + \left(R_{N_k} - \sum_{j=N_k}^{M_k-1} P_j \right) \right\} \\ &\leqq \lambda_d \left\{ R_{n_k} - \sum_{j=M_k}^{n_k-1} P_j \right\} \leqq 0, \qquad k = 1, 2, \cdots \end{aligned}$$

so that (1.1) is oscillatory by Theorem 2.2.

From Corollary 2.3, more particularly, we have Corollary 2.4.

COROLLARY 2.4. *If for any* N *there exists an* $n \geqq N + 1$ *such that*

$$\lambda_d \left(R_n - \sum_{j=N}^{n-1} P_j \right) \leqq 0 \tag{2.7}$$

and

$$\lim_{n \to \infty} \lambda_d \left(\sum_{j=0}^{n} P_j \right) = \infty, \tag{2.8}$$

then (1.1) *is oscillatory.*

Proof. For any integer N_k by (2.8) we can find an $M_k \geqq N_k + 1$ so that (2.5) holds. Then it follows from (2.7) that there is an $n_k \geqq M_k + 1$ so that (2.6) holds. Therefore, (1.1) is oscillatory by Corollary 2.3.

Remark. In the scalar case, Theorem 2.2 and Corollary 2.4 become Theorem 2.9 in [5] and Lemma 3.5 in [9], respectively.

Example 1. Let $d = 2$. For any integer $k \geqq 3$, we let $R_{k^2} = R_{k^2+9} = I$ (2×2 identity matrix), let $P_n = \operatorname{diag}(p_n, q_n)$ with $p_n = \frac{1}{4}$ and q_n arbitrary real numbers for $n = k^2, \cdots, k^2 + 8$, and let R_n, P_n be arbitrary otherwise. Then (1.1) is oscillatory since condition (2.4) in Theorem 2.2 holds for $N = k^2$, $n = k^2 + 9$. Note that (2.8) may not hold if we define P_n appropriately for $k^2 + 9 \leqq n \leqq (k+1)^2 - 1$.

We will denote by $\mathscr{C}$ the set of all Hermitian matrix sequences $f = \{f_n\}_{n=0}^{\infty}$ with the property that

$$\lim_{n \to \infty} \sum_{j=0}^{n} f_j = \sum_{j=0}^{\infty} f_j \quad \text{exists (finite).}$$

Clearly, if $f \in \mathscr{C}$ then $F_n = \sum_{j=n}^{\infty} f_j$ is Hermitian for all $n \geqq 0$. If f is a sequence of Hermitian matrices with $f_n \geqq 0$, $n \geqq 0$, then

$$f \in \mathscr{C} \Leftrightarrow \sum_{n=0}^{\infty} \operatorname{tr} f_n < \infty \Leftrightarrow \sum_{n=0}^{\infty} \|f_n\| < \infty. \tag{2.9}$$

Set $a_n = \lambda_d(R_n)$, $A_n = \lambda_1(R_n)$, $n \geqq 0$. Then

$$0 < a_n I \leqq R_n \leqq A_n I, \qquad 0 < A_n^{-1} I \leqq R_n^{-1} \leqq a_n^{-1} I. \tag{2.10}$$

We will also denote by $\mathscr{F}$ the set of all sequences of real numbers $b = \{b_n\}_{n=0}^{\infty}$ with $0 \leqq b_n \leqq 1$ and

$$\sum_{n=0}^{+\infty} b_n = +\infty. \tag{2.11}$$

It will be convenient to set

$$B_n = \sum_{j=0}^{n} b_j, \qquad B_{n,m} = \sum_{j=m}^{n} b_j.$$

Whenever we write $(B_{n,m})^{-1}$, we will take $n \geqq m$ sufficiently large so that $B_{n,m} > 0$. Obviously, we have

$$\lim_{n\to\infty} B_n^{-1} B_{n,N} = 1 \quad \text{for any fixed } N \geqq 0. \tag{2.12}$$

We also have the following lemma.

LEMMA 2.5. *Let $f = \{f_n\}_{n=0}^{\infty}$ be a sequence either of real numbers or of $d \times d$ matrices. If $\lim_{n\to\infty} f_n = C$ (where C is either a real number (or $\pm\infty$) or a $d \times d$ matrix), then*

$$\lim_{n\to\infty} (B_{n,N})^{-1} \sum_{k=N}^{n} b_k f_k = C \quad \text{for } N \geqq 0. \tag{2.13}$$

In the latter case, if C is a $d \times d$ matrix, then (2.13) *is to be interpreted componentwise.*

That is, "averaging" with respect to the class $\mathscr{F}$ preserves limits.

We introduce the following conditions, which will be needed in the results to follow:

(A1) The sequence A/a is bounded, i.e., there is a $K > 0$ such that $A_n/a_n \leqq K$, $n \geqq 0$, and there exist $b \in \mathscr{F}$ and $M > 0$ such that

$$B_n^{-3/2} \sum_{j=0}^{n} b_j A_{j+1} \leqq M \quad \text{for all large } n.$$

(A2) The sequence A/a is bounded and there exist $b \in \mathscr{F}$ and $M > 0$ such that

$$B_n^{-1} \sum_{j=0}^{n} b_j A_{j+1} \leqq M \quad \text{for all large } n.$$

(A3) The sequences A/a and A are bounded.

Clearly, we have (A3)$\Rightarrow$(A2)$\Rightarrow$(A1).

THEOREM 2.6. *If R satisfies* (A1) *and equation* (1.1) *is nonoscillatory, then the following are equivalent*:

(i) $$\lim_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} P_j = C, \tag{2.14}$$

where C is a constant Hermitian matrix and may depend on $b \in \mathscr{F}$;

(ii) $$\liminf_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j > -\infty; \tag{2.15}$$

(iii) *For any prepared solution* Y *of* (1.1) *with* $Y_{n+1}^* R_n Y_n > 0$ *for* $n \geqq N$ *for some* $N \geqq 0$, *the sequence*

$$Z(Z+R)^{-1}Z \in \mathscr{C}, \tag{2.16}$$

where Z *is given by* (1.8).

Proof. Obviously (i)$\Rightarrow$(ii).

(ii)$\Rightarrow$(iii). Suppose not, and let $\rho_n \equiv Z_n(Z_n+R_n)^{-1}Z_n$. Then since $\rho_n \geqq 0$ we have

$$\sum_{n=N}^{\infty} \operatorname{tr} \rho_n = +\infty. \tag{2.17}$$

From (1.9) we have

$$-Z_{n+1} = \sum_{j=N}^{n} Z_j(Z_j+R_j)^{-1}Z_j + \sum_{j=N}^{n} P_j - Z_N \tag{2.18}$$

and hence

$$\begin{aligned}(B_{n,N})^{-1} \sum_{k=N}^{n} b_k \sum_{j=N}^{k} \operatorname{tr} \rho_j + (B_{n,N})^{-1} \sum_{k=N}^{n} b_k \sum_{j=N}^{k} \operatorname{tr} P_j - \operatorname{tr} Z_N \\ = (B_{n,N})^{-1} \sum_{k=N}^{n} b_k \operatorname{tr}(-Z_{k+1}).\end{aligned} \tag{2.19}$$

Now by Lemma 2.5, (2.17) implies that the first term on the left side of (2.19) $\to \infty$ as $n \to \infty$. It follows from (ii) and (2.19) that

$$\lim_{n\to\infty} (B_{n,N})^{-1} \sum_{k=N}^{n} b_k \operatorname{tr}(-Z_{k+1}) = +\infty$$

and hence

$$\lim_{n\to\infty} (B_{n,N})^{-1} \sum_{k=N}^{n} b_k \|Z_{k+1}\| = +\infty. \tag{2.20}$$

Dividing (2.19) by $(B_{n,N})^{1/2}$, in view of (ii), condition (A1), and the fact that $-Z_{k+1} < R_{k+1}$, we have for all large n

$$(B_{n,N})^{-3/2} \sum_{k=N}^{n} b_k \sum_{j=N}^{k} \operatorname{tr} \rho_j \leqq M_1 < \infty.$$

For any fixed $n \geqq N+1$, choose $m > n$ so that $B_{n,N} \leqq (B_m - B_n) \leqq 2B_{n,N}$ and hence $B_{m,N} \leqq 3B_{n,N}$. Then we have

$$\begin{aligned}(B_{n,N})^{-1/2} \sum_{j=N}^{n} \operatorname{tr} \rho_j &\leqq (B_{n,N})^{-3/2}(B_m - B_n) \sum_{j=N}^{n} \operatorname{tr} \rho_j \\ &\leqq (B_{n,N})^{-3/2} \sum_{k=N}^{m} \left(b_k \sum_{j=N}^{k} \operatorname{tr} \rho_j \right) \\ &\leqq 3^{3/2}(B_{m,N})^{-3/2} \sum_{k=N}^{m} \left(b_k \sum_{j=N}^{k} \operatorname{tr} \rho_j \right) \leqq 3^{3/2} M_1 < \infty.\end{aligned} \tag{2.21}$$

On the other hand, since

$$\begin{aligned}\rho_n &= Z_n(Z_n+R_n)^{-1}Z_n = Z_n(R_n^{-1}Z_n + I)^{-1}R_n^{-1}Z_n \\ &= Z_n R_n^{-1} Z_n (I + R_n^{-1}Z_n)^{-1}\end{aligned}$$

and $A_n^{-1}Z_n^2 \leqq Z_n R_n^{-1} Z_n = \rho_n(I + R_n^{-1}Z_n)$, we have

$$\begin{aligned}\|Z_n\|^2 = \|Z_n^2\| &\leqq A_n \|\rho_n\| + \frac{A_n}{a_n}\|\rho_n\|\,\|Z_n\| \\ &= \frac{A_n}{a_n}\|\rho_n\|(a_n + \|Z_N\|) \\ &\leqq \frac{2A_n}{a_n}\|\rho_n\| \max(a_n, \|Z_n\|)\end{aligned}$$

and hence

$$\begin{aligned}\|Z_n\| &\leqq \max\left(\frac{2A_n}{a_n}\|\rho_n\|, (2\|\rho_n\|A_n)^{1/2}\right) \\ &\leqq 2K\|\rho_n\| + (2\|\rho_n\|A_n)^{1/2}.\end{aligned} \tag{2.22}$$

By the Schwarz inequality we obtain

$$\begin{aligned}(B_{n,N})^{-1}\sum_{k=N}^{n} b_k\|Z_{k+1}\| &\leqq 2K(B_{n,N})^{-1}\sum_{k=N}^{n} b_k\|\rho_{k+1}\| \\ &\quad + \sqrt{2}\,(B_{n,N})^{-1}\sum_{k=N}^{n} b_k(\|\rho_{k+1}\|A_{k+1})^{1/2} \\ &\leqq 2K(B_{n,N})^{-1}\sum_{k=N}^{n}\|\rho_{k+1}\| \\ &\quad + \sqrt{2}\,(B_{n,N})^{-1}\left(\sum_{k=N}^{n}\|\rho_{k+1}\|\right)^{1/2}\left(\sum_{k=N}^{n} b_k A_{k+1}\right)^{1/2}.\end{aligned} \tag{2.23}$$

Now from (2.23), (2.21), and (A1) we see that

$$(B_{n,N})^{-1}\sum_{k=N}^{n} b_k\|Z_{k+1}\|$$

is bounded above, which contradicts (2.20). This shows that (ii)$\Rightarrow$(iii).

(iii)$\Rightarrow$(i). From condition (iii) we have

$$\sum_{k=N}^{\infty}\|\rho_{k+1}\| < \infty. \tag{2.24}$$

Hence, from (2.23), (2.24), and (A1) it follows that

$$\lim_{n\to\infty}(B_{n,N})^{-1}\sum_{k=N}^{n} b_k\|Z_{k+1}\| = 0. \tag{2.25}$$

Since

$$(B_{n,N})^{-1}\sum_{k=N}^{n} b_k\sum_{j=N}^{k} P_j = Z_n + (B_{n,N})^{-1}\sum_{k=N}^{n} b_k(-Z_{k+1}) - (B_{n,N})^{-1}\sum_{k=N}^{n} b_k\sum_{j=N}^{k}\rho_j \tag{2.26}$$

and each term on the right side of (2.26) is convergent as $n\to\infty$, we obtain (i) by letting $n\to\infty$ in (2.26). This completes the proof of the theorem.

COROLLARY 2.7. *Suppose that R satisfies* (A1) *and that P satisfies condition* (2.15). *If*

$$\lim_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} P_j \quad \textit{does not exist (finite)}, \tag{2.27}$$

then (1.1) *is oscillatory.*

Remark. Thus, if there exists $b \in \mathcal{F}$ such that conditions (A1), (2.15), and (2.27) are satisfied, then we conclude that (1.1) is oscillatory. It is easy to give conditions under which the limit in (2.27) does not exist. For example, under the assumption that (A1) and (2.15) hold, (1.1) is oscillatory if either of the following conditions holds:

$$\limsup_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j = +\infty \tag{2.28}$$

or

$$\liminf_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j < \limsup_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j. \tag{2.29}$$

We next introduce a Riccati summation equation. Suppose that (A1) and condition (2.15) hold and that (1.1) is nonoscillatory. Since

$$B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} P_j = B_n^{-1} \sum_{k=0}^{N-1} b_k \sum_{j=0}^{k} P_j + B_n^{-1} B_{n,N} \left(\sum_{j=0}^{N-1} P_j \right) + (B_n^{-1} B_{n,N}) B_{n,N}^{-1} \sum_{k=N}^{n} b_k \sum_{j=N}^{k} P_j$$

for $n > N \geqq 0$, we have after some computation

$$\lim_{n\to\infty} B_{n,N}^{-1} \sum_{k=N}^{n} b_k \sum_{j=N}^{k} P_j = C - \sum_{j=0}^{N-1} P_j \tag{2.30}$$

(using $B_n^{-1} B_{n,N} \to 1$ as $n \to \infty$), where C is defined as in (2.14).

THEOREM 2.8. *Let* (A1) *and condition* (2.15) *hold for some* $b \in \mathcal{F}$. *Then* (1.1) *is nonoscillatory if and only if there exists a sequence Z of Hermitian matrices with* $Z_n > -R_n$ *for* $n \geqq N$ *for some* $N \geqq 0$ *satisfying*

$$Z_n = C - \sum_{j=0}^{n-1} P_j + \sum_{j=n}^{\infty} Z_j (Z_j + R_j)^{-1} Z_j, \qquad n \geqq N, \tag{2.31}$$

where C is defined in (2.14) *and may depend on b.*

Proof. Suppose first that (1.1) is nonoscillatory. By Theorem 2.6 there exists a sequence Z of Hermitian matrices with $Z_n > -R_n$ for $n \geqq N$ satisfying (2.26). Letting $n \to \infty$ in (2.26), using (2.30), and replacing N by n, we obtain (2.31). Conversely, suppose that Z is the Hermitian matrix sequence in the theorem. From (2.31) we have

$$\Delta Z_n = -P_n - Z_n (Z_n + R_n)^{-1} Z_n$$

and so (1.1) is nonoscillatory by Theorem 2.1.

THEOREM 2.9. *Let* (A2) *hold for some* $b \in \mathcal{F}$. *If* (1.1) *is nonoscillatory then the following statements are equivalent*:

(i) $\lim_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j = -\infty$;

(ii) $\liminf_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j = -\infty$;

(iii) *There exists a nonoscillatory solution Y of* (1.1) *with* $Y_{n+1}^* R_n Y_N > 0$ *for* $n \geqq N$ *for some* $N \geqq 0$ *such that the sequence* $Z_n = R_n (\Delta Y_n) Y_n^{-1}$, $n \geqq N$, *satisfies*

$$\sum_{j=N}^{\infty} \operatorname{tr} (Z_j (Z_j + R_j)^{-1} Z_j) = +\infty.$$

Proof. Since (A2)$\Rightarrow$(A1), (ii) and (iii) are equivalent by virtue of Theorem 2.6. Since (i) clearly implies (ii) we need only show that (ii) and (iii)$\Rightarrow$(i). From (2.19)

we have

$$(B_{n,N})^{-1}\sum_{k=N}^{n} b_k \sum_{j=N}^{k} \operatorname{tr} P_j \leqq \operatorname{tr} Z_n + (B_{n,N})^{-1}\sum_{k=N}^{n} b_k \operatorname{tr} R_{k+1} - (B_{n,N})^{-1}\sum_{k=N}^{n} b_k \sum_{j=N}^{k} \operatorname{tr} \rho_j. \tag{2.32}$$

Since the right side of (2.32) tends to $-\infty$ as $n\to\infty$, it follows that (i) holds. This completes the proof.

COROLLARY 2.10. *If* (A2) *holds for some* $b\in\mathscr{F}$ *and*

$$-\infty = \liminf_{n\to\infty} B_n^{-1}\sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j < \limsup_{n\to\infty} B_n^{-1}\sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j,$$

then (1.1) *is oscillatory.*

THEOREM 2.11. *Let* (A3) *hold. If* (1.1) *is nonoscillatory, then the following statements are equivalent*:

(i) $P\in\mathscr{C}$;

(ii) $\lim_{n\to\infty} B_n^{-1}\sum_{k=0}^{n} b_k \sum_{j=0}^{k} P_j = C$ *for some* $b\in\mathscr{F}$, *where* C *is a constant matrix*;

(iii) *The sequence* P *satisfies condition* (2.15) *for some* $b\in\mathscr{F}$;

(iv) *For any nonoscillatory solution* Y *of* (1.1) *with* $Y_{n+1}^{*}R_nY_n>0$ *for* $n\geqq N$ *for some* $N\geqq 0$ *the sequence* $Z_n = R_n(\Delta Y_n)Y_n^{-1}$, $n\geqq N$, *satisfies* $Z(Z+R)^{-1}Z\in\mathscr{C}$.

Proof. Since (A3)$\Rightarrow$(A1), (iii) and (iv) are equivalent by Theorem 2.6. Obviously (i)$\Rightarrow$(ii)$\Rightarrow$(iii). We need only show that (iii) and (iv)$\Rightarrow$(i). Now from (2.18) it is sufficient to show that $Z_n\to 0$ as $n\to\infty$. But this is immediate from (2.22) because of the boundedness of the sequence $\{A_n\}$. This completes the proof.

We observe that if $P\in\mathscr{C}$, then the limit C in (ii) of Theorem 2.11 is independent of $b\in\mathscr{F}$ and $C=\sum_{j=0}^{\infty} P_j$. If we let

$$Q_n = \sum_{j=n}^{\infty} P_j,$$

then we can get an analogue of Theorem 2.8 under the assumption (A3) instead of (A1) as follows.

THEOREM 2.12. *Let* (A3) *and condition* (2.15) *hold for some* $b\in\mathscr{F}$. *Then* (1.1) *is nonoscillatory if and only if* $P\in\mathscr{C}$ *and there exists a sequence* Z *of Hermitian matrices with* $Z_n>-R_n$ *for* $n\geqq N$ *for some* $N\geqq 0$ *satisfying*

$$Z_n = Q_n + \sum_{j=n}^{\infty} Z_j(Z_j+R_j)^{-1}Z_j, \qquad n\geqq N. \tag{2.33}$$

We can also obtain a counterpart to Theorem 2.11.

THEOREM 2.13. *Let* (A3) *hold. If* (1.1) *is nonoscillatory then the following statements are equivalent*:

(i) $\lim_{n\to\infty}\lambda_d(\sum_{j=0}^{n} P_j)=-\infty$, $\limsup_{n\to\infty}\lambda_1(\sum_{j=0}^{n} P_j)<\infty$;

(ii) $\lim_{n\to\infty}\operatorname{tr}(\sum_{j=0}^{n} P_j)=-\infty$;

(iii) $\liminf_{n\to\infty} B_n^{-1}\sum_{k=0}^{n} b_k\sum_{j=0}^{k}\operatorname{tr} P_j=-\infty$ *for some* $b\in\mathscr{F}$;

(iv) *There exists a prepared solution* Y *of* (1.1) *with* $Y_{n+1}^{*}R_nY_n>0$ *for* $n\geqq N$ *for some* $N\geqq 0$ *such that*

$$\sum_{j=N}^{\infty}\operatorname{tr}(Z_j(Z_j+R_j)^{-1}Z_j)=\infty,$$

where $Z_n = R_n(\Delta Y_n)Y_n^{-1}>-R_n$ *for* $n\geqq N$.

Proof. Obviously (i)⇒(ii)⇒(iii). By Theorem 2.11 it follows that (iii) and (iv) are equivalent. If Z is the sequence in (iv), then for all $n \geqq N$ we have

$$\sum_{j=N}^{n} P_j = Z_N - Z_{n+1} - \sum_{j=N}^{n} Z_j(Z_j + R_j)^{-1} Z_j$$

$$< Z_N + R_{n+1} - \sum_{j=N}^{n} Z_j(Z_j + R_j)^{-1} Z_j.$$

Hence

$$\operatorname{tr}\left(\sum_{j=N}^{n} P_j\right) \leqq \operatorname{tr} Z_N + M_1 - \sum_{j=N}^{n} \operatorname{tr}\left(Z_j(Z_j + R_j)^{-1} Z_j\right) \to -\infty$$

as $n \to \infty$, i.e., (iv)⇒(ii). By the convexity of the functional $\lambda_1(\cdot)$, we also have

$$\lambda_1\left(\sum_{j=N}^{n} P_j\right) \leqq \lambda_1(Z_N) + \lambda_1(R_{n+1}) + \lambda_1\left(-\sum_{j=N}^{n} Z_j(Z_j + R_j)^{-1} Z_j\right) \leqq \lambda_1(Z_N) + M_1,$$

i.e., (iv) and (ii)⇒(i). This proves the theorem.

Remark. Theorems 2.11 and 2.13 show the significant dependence of the oscillation of (1.1) on the divergence of the sequence P under (A3). The analogue of Theorem 2.11 for the differential equation (1.2) is only partially true. For example, if we let $R(t) = I$ and assume

$$\liminf_{t\to\infty} \frac{1}{t} \int_0^t \int_0^s \operatorname{tr} P(r)\, dr\, ds > -\infty,$$

then the nonoscillation of (1.2) does not imply the convergence of $\int_0^\infty P(t)\, dt$.

The following oscillation criteria are immediate from Theorems 2.11 and 2.13.

COROLLARY 2.14. *Let* (A3) *hold and assume that P satisfies* (2.15) *for some* $b \in \mathcal{F}$. *If* $P \notin \mathcal{C}$, *then* (1.1) *is oscillatory. In particular,* (1.1) *is oscillatory in case either*

$$\limsup_{n\to\infty} \lambda_1\left(\sum_{j=0}^{n} P_j\right) = +\infty, \tag{2.34}$$

or

$$\liminf_{n\to\infty} \lambda_k\left(\sum_{j=0}^{n} P_j\right) < \limsup_{n\to\infty} \lambda_k\left(\sum_{j=0}^{n} P_j\right) \quad \text{for some } 1 \leqq k \leqq d. \tag{2.35}$$

COROLLARY 2.15. *Let* (A3) *hold. If for some* $b \in \mathcal{F}$

$$\liminf_{n\to\infty} B_n^{-1} \sum_{k=0}^{n} b_k \sum_{j=0}^{k} \operatorname{tr} P_j = -\infty,$$

then (1.1) *is oscillatory in case either* (2.34) *holds or*

$$\limsup_{n\to\infty} \sum_{j=0}^{n} \operatorname{tr} P_j > -\infty. \tag{2.36}$$

Combining Corollaries 2.14 and 2.15, we obtain Corollary 2.16.

COROLLARY 2.16. *If* (A3) *and* (2.34) *hold, then* (1.1) *is oscillatory.*

Remark. Corollary 2.16 shows that the discrete matrix analogue of a recent result of Kaper and Kwong [8] and of Byers, Harris, and Kwong [4] for the continuous case (eq. (1.2)) is valid. The criteria in Corollaries 2.14 and 2.15 may also be considered analogues of some of the results of [3] for (1.2).

Example 2. Let $d=2$, $R_n=I$ and let $P_n=\operatorname{diag}(p_n, q_n)$ with $p_n+q_n=0$, $p_n, q_n<2$ for all n, and

$$\liminf_{n\to\infty}\sum_{j=0}^{n} p_j<\limsup_{n\to\infty}\sum_{j=0}^{n} p_j.$$

Then (1.1) is oscillatory by Corollary 2.14.

Example 3. Let $d=2$, $R_n=I$, and $P_n=\operatorname{diag}(p_n, q_n)$ with $p_n=1/n$ if n is prime and $p_n=0$ otherwise. If

$$\limsup_{n\to\infty}\sum_{j=0}^{n}(p_j+q_j)>-\infty,$$

then (1.1) is oscillatory from the following result.

COROLLARY 2.17. *Let* (A3) *hold. If* $P\notin\mathscr{C}$ *and* (2.36) *holds, then* (1.1) *is oscillatory.*

Proof. From (2.36) we can find a sequence of integers n_r with $n_r\to\infty$ as $r\to\infty$ such that

$$\lim_{r\to\infty}\sum_{j=0}^{n_r}\operatorname{tr} P_j=C>-\infty.$$

Let $b_n=1$ if $n=n_r$, $r=1,2,\cdots$ and $b_n=0$ otherwise. Then

$$B_{n_r}^{-1}\sum_{k=0}^{n_r} b_k\sum_{j=0}^{k}\operatorname{tr} P_j=\frac{1}{r}\sum_{k=1}^{r}\sum_{j=0}^{n_k}\operatorname{tr} P_j\to C\quad\text{as } r\to\infty.$$

Thus, (2.15) is satisfied and the oscillation of (1,1) follows from Corollary 2.14.

More sophisticated examples may also be given. We leave this to the interested reader.

Acknowledgment. The authors are indebted to the referees for their detailed comments.

REFERENCES

[1] C. D. AHLBRANDT AND J. W. HOOKER, *Recessive solutions of symmetric three term recurrence relations*, in C.M.S. Conference Proc., Vol. 8, Oscillation, Bifurcation and Chaos, American Mathematical Society, Providence, RI, 1987, pp. 3-42.

[2] ———, *Riccati matrix difference equations and disconjugacy of discrete linear systems*, SIAM J. Math. Anal., 19 (1988), 1183-1197.

[3] G. J. BUTLER, L. H. ERBE, AND A. B. MINGARELLI, *Riccati techniques and variational principles in oscillation theory for linear systems*, Trans. Amer. Math. Soc., 303 (1987), pp. 263-282.

[4] R. BYERS, B. J. HARRIS, AND M. K. KWONG, *Weighted means and oscillation conditions for second order matrix differential equations*, J. Differential Equations, 61 (1986), pp. 164-177.

[5] S. CHEN AND L. H. ERBE, *Riccati techniques and discrete oscillations*, J. Math. Anal. Appl., to appear.

[6] D. B. HINTON AND R. T. LEWIS, *Spectral analysis of second order difference equations*, J. Math. Anal. Appl., 63 (1978), pp. 421-438.

[7] J. W. HOOKER, M. K. KWONG, AND W. T. PATULA, *Oscillatory second order linear difference equations and Riccati equations*, SIAM J. Math. Anal., 18 (1987), pp. 54-63.

[8] H. G. KAPER AND M. K. KWONG, *Oscillation theory for linear second-order differential systems*, in CMS Conference Proc., Vol. 8, Oscillation, Bifurcation and Chaos, American Mathematical Society, Providence, RI, 1987, pp. 187-198.

[9] M. K. KWONG, J. W. HOOKER, AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations*, II., J. Math. Anal. Appl., 107 (1985), pp. 182-196.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 950-965, July 1989

011

CONDITIONAL ANALYTIC FEYNMAN INTEGRALS AND A RELATED SCHRÖDINGER INTEGRAL EQUATION*

DONG MYUNG CHUNG† AND DAVID SKOUG‡

Abstract. The concept of a conditional Feynman integral of a function F given a function X is introduced, and its existence is established for various functions. Then the conditional Feynman integral is used to derive an integral equation that is formally equivalent to the Schrödinger equation.

Key words. Wiener integral, conditional Wiener integral, Feynman integral, conditional Feynman integral

AMS(MOS) subject classification. 28C20

1. Introduction. In this paper we define the concept of a conditional analytic Feynman integral of a function F on Wiener space given a function X. In particular, for a certain choice of X we establish the existence of the conditional Feynman integral for all functions F in the Banach algebra S introduced by Cameron and Storvick [3]. In [15] Johnson has shown that S is isometrically isomorphic to the class $\mathscr{F}(H)$ of Fresnel integrable functions as given by Albeverio and Høegh-Krohn [1]. The Banach algebras $\mathscr{F}(H)$ and S and related theories have been studied quite extensively; references include [3]-[7], [9]-[11], [14], [15], [17]-[21].

To define the conditional analytic Feynman integral we use the conditional Wiener integral as given by Yeh in [23], [24] and studied further in [8], [13], [22], [25]. In [24] Yeh used the conditional Wiener integral to derive the Kac-Feynman formula. Here we use the conditional Feynman integral to derive an integral equation formally equivalent to the Schrödinger equation. We also use the conditional Feynman integral to provide a fundamental solution to the Schrödinger equation.

2. Definitions and preliminaries. Let ν be a positive integer and let $C_0^\nu[0, t]$ denote ν-dimensional Wiener space, that is the space of $\mathbb{R}^\nu$-valued continuous functions $\vec{x}$ on $[0, t]$ such that $\vec{x}(0) = \vec{0}$. Let $\mathscr{M}^\nu$ denote the class of all Wiener measurable subsets of $C_0^\nu[0, t]$ and let m^ν denote ν-dimensional Wiener measure. $(C_0^\nu[0, t], \mathscr{M}^\nu, m^\nu)$ is a complete measure space. In case $\nu = 1$, we delete the one and simply write $(C_0[0, t], \mathscr{M}, m)$. Of course, $C_0^\nu[0, t] = C_0[0, t] \times \cdots \times C_0[0, t]$ (ν times). We denote the Wiener integral of a function F by

$$\int_{C_0^\nu[0,t]} F(\vec{x})\, dm^\nu(\vec{x}).$$

A subset E of $C_0^\nu[0, t]$ is said to be scale-invariant measurable [12], [16] provided ρE is Wiener measurable for each $\rho > 0$, and a scale-invariant measurable set N is said to be scale-invariant null provided $m^\nu(\rho N) = 0$ for each $\rho > 0$. A property that holds except on a scale-invariant null set is said to hold scale-invariant almost everywhere (s-a.e.).

Next we give Yeh's definition of the conditional Wiener integral from [24].

* Received by the editors July 20, 1987; accepted for publication (in revised form) October 4, 1988.

† Sogang University, Seoul 121, Korea. The work of this author was supported by the Basic Science Institute Research Program and the Ministry of Education of Korea.

‡ Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588-0323.

DEFINITION 1. Let X be an $\mathbb{R}^\nu$-valued Wiener measurable function on $C_0^\nu[0, t]$ and let F be a complex-valued Wiener integrable function on $C_0^\nu[0, t]$. Let P_X be the probability distribution of X, i.e., for all $B \in \mathcal{B}^\nu$, the Borel sets in $\mathbb{R}^\nu$, $P_X(B) = m^\nu(X^{-1}(B))$. The conditional Wiener integral of F given X is by definition the equivalence class of Borel measurable and P_X-integrable functions ψ on $\mathbb{R}^\nu$, modulo null functions on $(\mathbb{R}^\nu, \mathcal{B}^\nu, P_X)$, such that for all $B \in \mathcal{B}^\nu$,

$$\int_{X^{-1}(B)} F(\vec{x})\, dm^\nu(\vec{x}) = \int_B \psi(\vec{\eta})\, dP_X(\vec{\eta}).$$

By the Radon-Nikodym theorem such a function ψ exists and is determined up to a null function on $(\mathbb{R}^\nu, \mathcal{B}^\nu, P_X)$. We let $E(F|X)$ denote a representative of the equivalence class and so for all $B \in \mathcal{B}^\nu$,

$$\int_{X^{-1}(B)} F(\vec{x})\, dm^\nu(\vec{x}) = \int_B E(F|X)(\vec{\eta})\, dP_X(\vec{\eta}).$$

We are now ready to define the conditional analytic Feynman integral of a function F given X.

DEFINITION 2. Let X be an $\mathbb{R}^\nu$-valued scale-invariant measurable function on $C_0^\nu[0, t]$ and let F be a scale-invariant measurable function on $C_0^\nu[0, t]$ such that the Wiener integral

$$\int_{C_0^\nu[0,t]} F(\lambda^{-1/2}\vec{x})\, dm^\nu(\vec{x})$$

exists as a finite number for all $\lambda > 0$. For $\lambda > 0$ let

$$J_\lambda(\vec{\eta}) = E(F(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot))(\vec{\eta})$$

denote the conditional Wiener integral of $F(\lambda^{-1/2}\cdot)$ given $X(\lambda^{-1/2}\cdot)$. If for almost every $\vec{\eta} \in \mathbb{R}^\nu$, there exists a function $J_\lambda^*(\vec{\eta})$, analytic in λ on $\mathbb{C}^+ \equiv \{\lambda \in \mathbb{C}: \operatorname{Re} \lambda > 0\}$ such that $J_\lambda^*(\vec{\eta}) = J_\lambda(\vec{\eta})$ for all $\lambda > 0$, then J_λ^* is defined to be the conditional analytic Wiener integral of F given X with parameter λ and for $\lambda \in \mathbb{C}^+$ we write

$$E^{\mathrm{anw}_\lambda}(F|X)(\vec{\eta}) = J_\lambda^*(\vec{\eta}).$$

If for fixed real $q \neq 0$, the limit

$$\lim_{\lambda \to -iq} E^{\mathrm{anw}_\lambda}(F|X)(\vec{\eta})$$

exists for almost every $\vec{\eta} \in \mathbb{R}^\nu$ where λ approaches $-iq$ through $\mathbb{C}^+$, we will denote the value of this limit by $E^{\mathrm{anf}_q}(F|X)$ and call it the conditional analytic Feynman integral of F given X with parameter q.

Remark 1. The notation $E^{\mathrm{anw}_\lambda}(F|X)$ does not mean "conditional expectation with respect to a probability measure" but rather an extension of such a conditional expectation.

Remark 2. In [24] Yeh (in this paper he worked with $\nu = 1$ but clearly his results in § 3 hold for general ν) always chose X to be the function $X(\vec{x}) = \vec{x}(t)$ and in that case

$$\frac{dP_X(\vec{\eta})}{d\vec{\eta}} = (2\pi t)^{-\nu/2} \exp\left\{-\frac{\|\vec{\eta}\|^2}{2t}\right\}.$$

In addition, when we use his inversion formulas [23, Thm. 2] and [24, Thm. 3], a version of $E(F|X)(\vec{\eta})$ is given for all $\vec{\eta} \in \mathbb{R}^\nu$ by the formula

$$(1)\quad E(F|X)(\vec{\eta}) = \left(\frac{dP_X(\vec{\eta})}{d\vec{\eta}}\right)^{-1} (2\pi)^{-\nu} \int_{\mathbb{R}^\nu} e^{-i\langle \vec{U}, \vec{\eta}\rangle} \int_{C_0^\nu[0,t]} e^{i\langle \vec{U}, \vec{x}(t)\rangle} F(\vec{x})\, dm^\nu(\vec{x})\, d\vec{U}$$

provided that the Wiener integral

$$\int_{C_0^\nu[0,t]} e^{i\langle \vec{U}, \vec{x}(t)\rangle} F(\vec{x})\, dm^\nu(\vec{x}) \tag{2}$$

is an integrable function of $\vec{U}$ on $\mathbb{R}^\nu$.

Remark 3. In [22], Park and Skoug have shown that if F is Borel measurable and Wiener integrable and if $X(\vec{x}) = \vec{x}(t)$, then the conditional Wiener integral $E(F|X)$ is given in terms of an ordinary Wiener integral by the formula

$$E(F|X)(\vec{\eta}) = \int_{C_0^\nu[0,t]} F\left(\vec{x}(\cdot) - \frac{\cdot}{t}\vec{x}(t) + \frac{\cdot}{t}\vec{\eta}\right) dm^\nu(\vec{x}). \tag{3}$$

Since all the functions F we consider in this paper are Borel measurable we can use either (1) or (3) to compute $E(F|X)$; one advantage of using (3) is that we do not need to show first that the expression in (2) is an integrable function of $\vec{U}$ on $\mathbb{R}^\nu$. It has also been pointed out in [22] that if F is only Wiener measurable rather than Borel measurable then the expression on the right-hand side of (3) is not necessarily a Borel measurable function of $\vec{\eta}$; however, in that case we can still choose a version of $E(F|X)$ that is equal almost everywhere to the right-hand side of (3).

We finish this section by stating the definitions [3] of the analytic Feynman integral and the Banach algebra $S(\nu)$.

Let F be a $\mathbb{C}$-valued scale-invariant measurable function on $C_0^\nu[0, t]$ such that the Wiener integral

$$J(\lambda) = \int_{C_0^\nu[0,t]} F(\lambda^{-1/2}\vec{x})\, dm^\nu(\vec{x})$$

exists as a finite number for all $\lambda > 0$. If there exists a function $J^*(\lambda)$ analytic in $\mathbb{C}^+$ such that $J^*(\lambda) = J(\lambda)$ for all $\lambda > 0$, then $J^*(\lambda)$ is defined to be the analytic Wiener integral of F over $C_0^\nu[0, t]$ with parameter λ, and for $\lambda \in \mathbb{C}^+$ we write

$$E^{\mathrm{anw}_\lambda}(F) = J^*(\lambda).$$

Let $q \neq 0$ be a real number and let F be a function such that $E^{\mathrm{anw}_\lambda}(F)$ exists for all $\lambda \in \mathbb{C}^+$. If the following limit exists, we call it the analytic Feynman integral of F with parameter q and we write

$$E^{\mathrm{anf}_q}(F) = \lim_{\lambda \to -iq} E^{\mathrm{anw}_\lambda}(F)$$

where λ approaches $-iq$ through $\mathbb{C}^+$.

The Banach algebra $S(\nu)$ consists of functions on $C_0^\nu[0, t]$ expressible in the form

$$F(\vec{x}) = \int_{L_2^\nu[0,t]} \exp\left\{ i \sum_{j=1}^{\nu} \int_0^t v_j(s)\, \tilde{d}x_j(s) \right\} d\sigma(\vec{V}) \tag{4}$$

for s-a.e. $\vec{x} = (x_1, \cdots, x_\nu)$ in $C_0^\nu[0, t]$ where σ is an element of $M(L_2^\nu[0, t])$, the space of $\mathbb{C}$-valued, countably additive Borel measures on $L_2^\nu[0, t]$, and the integrals $\int_0^t v_j(s)\, \tilde{d}x_j(s)$ are Paley-Wiener-Zygmund (PWZ) integrals. In addition, Cameron and Storvick [3, Thm. 5.1] show that for F given by (4)

$$E^{\mathrm{anw}_\lambda}(F) = \int_{L_2^\nu[0,t]} \exp\left\{ -\frac{1}{2\lambda} \sum_{j=1}^{\nu} \|v_j\|^2 \right\} d\sigma(\vec{V}), \qquad \lambda \in \mathbb{C}^+$$

and

$$E^{\mathrm{anf}_q}(F) = \int_{L_2^\nu[0,t]} \exp\left\{ -\frac{i}{2q} \sum_{j=1}^{\nu} \|v_j\|^2 \right\} d\sigma(\vec{V}) \tag{5}$$

for each real $q \neq 0$.

Finally we state the following well-known integration formula:

$$\int_{\mathbb{R}^\nu} \exp\left\{-\frac{b}{2}\|\vec{\eta}\|^2 + i\langle\vec{\eta},\vec{\xi}\rangle\right\} d\vec{\eta} = \left(\frac{2\pi}{b}\right)^{\nu/2} \exp\left\{-\frac{1}{2b}\|\vec{\xi}\|^2\right\}, \qquad \text{Re } b > 0,$$

which we use several times in this paper.

3. Conditional analytic Feynman integrals.

THEOREM 1. *Let $F \in S(\nu)$ be given by* (4) *and let $X(\vec{x}) = \vec{x}(t)$. Then for all $\lambda \in \mathbb{C}^+$, the conditional analytic Wiener integral $E^{\text{anw}_\lambda}(F|X)$ exists and for all $\vec{\eta} \in \mathbb{R}^\nu$ is given by the formula*

$$E^{\text{anw}_\lambda}(F|X)(\vec{\eta}) = \int_{L_2^\nu[0,t]} \exp\left\{-\frac{1}{2\lambda t}\sum_{j=1}^{\nu}[t\|v_j\|^2 - b_j^2] + \frac{i}{t}\langle\vec{\eta},\vec{B}\rangle\right\} d\sigma(\vec{V}) \tag{6}$$

where $\vec{B} = (b_1, \cdots, b_\nu) = (\int_0^t v_1(s)\,ds, \cdots, \int_0^t v_\nu(s)\,ds)$. Furthermore, the conditional analytic Feynman integral $E^{\text{anf}_q}(F|X)$ exists for all real $q \neq 0$ and for all $\vec{\eta} \in \mathbb{R}^\nu$ is given by the formula

$$E^{\text{anf}_q}(F|X)(\vec{\eta}) = \int_{L_2^\nu[0,t]} \exp\left\{-\frac{i}{2qt}\sum_{j=1}^{\nu}[t\|v_j\|^2 - b_j^2] + \frac{i}{t}\langle\vec{\eta},\vec{B}\rangle\right\} d\sigma(\vec{V}). \tag{7}$$

Proof. Using (3), the Fubini theorem, and a fundamental Wiener integration formula involving PWZ integrals, we obtain that for $\lambda > 0$ and all $\vec{\eta} \in \mathbb{R}^\nu$

$$\begin{aligned}
&E(F(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot))(\vec{\eta}) \\
&= \int_{C_0^\nu[0,t]} \int_{L_2^\nu[0,t]} \exp\left\{i\sum_{j=1}^{\nu}\int_0^t v_j(s)\,\tilde{d}\left[\lambda^{-1/2}x_j(s) - \lambda^{-1/2}\frac{s}{t}x_j(t) + \frac{s}{t}\eta_j\right]\right\} d\sigma(\vec{V})\,dm^\nu(\vec{x}) \\
&= \int_{L_2^\nu[0,t]} \int_{C_0^\nu[0,t]} \exp\left\{\frac{i}{\sqrt{\lambda}}\sum_{j=1}^{\nu}\left[\int_0^t v_j(s)\,\tilde{d}x_j(s) - \frac{x_j(t)}{t}\int_0^t v_j(s)\,ds\right] + \frac{i}{t}\sum_{j=1}^{\nu}\eta_j\int_0^t v_j(s)\,ds\right\} dm^\nu(\vec{x})\,d\sigma(\vec{V}) \qquad (8) \\
&= \int_{L_2^\nu[0,t]} \exp\left\{\frac{i}{t}\langle\vec{\eta},\vec{B}\rangle\right\} \int_{C_0^\nu[0,t]} \cdot \exp\left\{\frac{i}{\sqrt{\lambda}}\sum_{j=1}^{\nu}\int_0^t\left[v_j(s) - \frac{b_j}{t}\right]\tilde{d}x_j(s)\right\} dm^\nu(\vec{x})\,d\sigma(\vec{V}) \\
&= \int_{L_2^\nu[0,t]} \exp\left\{\frac{i}{t}\langle\vec{\eta},\vec{B}\rangle\right\} \exp\left\{-\frac{1}{2\lambda}\sum_{j=1}^{\nu}\int_0^t\left[v_j(s) - \frac{b_j}{t}\right]^2 ds\right\} d\sigma(\vec{V}) \\
&= \int_{L_2^\nu[0,t]} \exp\left\{-\frac{1}{2\lambda t}\sum_{j=1}^{\nu}\left[t\|v_j\|^2 - b_j^2\right] + \frac{i}{t}\langle\vec{\eta},\vec{B}\rangle\right\} d\sigma(\vec{V}).
\end{aligned}$$

But by the Cauchy-Schwarz inequality it follows that

$$b_j^2 = \left[\int_0^t v_j(s)\,ds\right]^2 \leqq \int_0^t 1^2\,ds \int_0^t v_j^2(s)\,ds = t\|v_j\|^2. \tag{9}$$

Hence, since $\sigma \in M(L_2^\nu[0,t])$, we see that the last expression on the right-hand side of (8) is an analytic function of λ throughout $\mathbb{C}^+$ and is a continuous function of λ for $\text{Re } \lambda \geqq 0$, $\lambda \neq 0$. Thus (see Definition 2 above) (6) and (7) are established, which concludes the proof of Theorem 1.

In our next theorem we show that if we multiply $E^{\text{anf}_q}(F|X)(\vec{\eta})$ by $(q/2\pi it)^{\nu/2}\exp\{(iq/2t)\|\vec{\eta}\|^2\}$, the analytic extension of the Radon-Nikodym derivative evaluated at $\lambda=-iq$, and then integrate over $\mathbb{R}^\nu$ we obtain the Feynman integral $E^{\text{anf}_q}(F)$. However, to do so we need a summation procedure; we will use the same one as in [19, p. 340]. Let

$$\overline{\int}_{\mathbb{R}^\nu} f(\vec{\eta})\,d\vec{\eta}=\lim_{A\to+\infty}\int_{\mathbb{R}^\nu} f(\vec{\eta})\exp\left\{-\frac{\|\vec{\eta}\|^2}{2A}\right\}d\vec{\eta} \tag{10}$$

whenever the expression on the right exists. Of course, if $f\in L_1(\mathbb{R}^\nu)$, it is clear using the Dominated Convergence Theorem that

$$\overline{\int}_{\mathbb{R}^\nu} f(\vec{\eta})\,d\vec{\eta}=\int_{\mathbb{R}^\nu} f(\vec{\eta})\,d\vec{\eta}.$$

THEOREM 2. *Let $F\in S(\nu)$ be given by* (4) *and let $X(\vec{x})=\vec{x}(t)$. Then for all $\lambda\in\mathbb{C}^+$*

$$\int_{\mathbb{R}^\nu}\left(\frac{\lambda}{2\pi t}\right)^{\nu/2}\exp\left\{-\frac{\lambda}{2t}|\vec{\eta}\|^2\right\}E^{\text{anw}_\lambda}(F|X)(\vec{\eta})\,d\vec{\eta}=E^{\text{anw}_\lambda}(F) \tag{11}$$

and for all real $q\neq 0$,

$$\overline{\int}_{\mathbb{R}^\nu}\left(\frac{q}{2\pi it}\right)^{\nu/2}\exp\left\{\frac{iq}{2t}\|\vec{\eta}\|^2\right\}E^{\text{anf}_q}(F|X)(\vec{\eta})\,d\vec{\eta}=E^{\text{anf}_q}(F). \tag{12}$$

Proof. We will establish (12); the proof of (11) is similar, but easier since the summation procedure is not needed. Let $q\neq 0$ be given. Then using (10), (7), the Fubini theorem, and (5), we obtain

$$\begin{aligned}
&\overline{\int}_{\mathbb{R}^\nu}\left(\frac{q}{2\pi it}\right)^{\nu/2}\exp\left\{\frac{iq}{2t}\|\vec{\eta}\|^2\right\}E^{\text{anf}_q}(F|X)(\vec{\eta})\,d\vec{\eta}\\
&\quad=\lim_{A\to+\infty}\int_{\mathbb{R}^\nu}\left(\frac{q}{2\pi it}\right)^{\nu/2}\exp\left\{\frac{1}{2}\left(\frac{iq}{t}-\frac{1}{A}\right)\|\vec{\eta}\|^2\right\}\\
&\qquad\cdot\int_{L_2^\nu[0,t]}\exp\left\{-\frac{i}{2qt}\sum_{j=1}^{\nu}[t\|v_j\|^2-b_j^2]+\frac{i}{t}\langle\vec{\eta},\vec{B}\rangle\right\}d\sigma(\vec{V})\,d\vec{\eta}\\
&\quad=\lim_{A\to+\infty}\int_{L_2^\nu[0,t]}\exp\left\{-\frac{i}{2qt}\sum_{j=1}^{\nu}[t\|v_j\|^2-b_j^2]\right\}\left(\frac{q}{2\pi it}\right)^{\nu/2}\\
&\qquad\cdot\int_{\mathbb{R}^\nu}\exp\left\{-\frac{1}{2}\left(\frac{t-Aiq}{At}\right)\|\vec{\eta}\|^2+i\left\langle\vec{\eta},\frac{\vec{B}}{t}\right\rangle\right\}d\vec{\eta}\,d\sigma(\vec{V})\\
&\quad=\lim_{A\to+\infty}\int_{L_2^\nu[0,t]}\exp\left\{-\frac{i}{2qt}\sum_{j=1}^{\nu}[t\|v_j\|^2-b_j^2]\right\}\left(\frac{q}{2\pi it}\right)^{\nu/2}\\
&\qquad\cdot\left(\frac{2\pi At}{t-Aiq}\right)^{\nu/2}\exp\left\{-\frac{A}{2(t-Aiq)t}\|\vec{B}\|^2\right\}d\sigma(\vec{V})\\
&\quad=\int_{L_2^\nu[0,t]}\exp\left\{-\frac{i}{2q}\sum_{j=1}^{\nu}\|v_j\|^2+\frac{i}{2qt}\|\vec{B}\|^2\right\}\exp\left\{-\frac{i}{2qt}\|\vec{B}\|^2\right\}d\sigma(\vec{V})\\
&\quad=\int_{L_2^\nu[0,t]}\exp\left\{-\frac{i}{2q}\sum_{j=1}^{\nu}\|v_j\|^2\right\}d\sigma(\vec{V})\\
&\quad=E^{\text{anf}_q}(F).
\end{aligned}$$

In our next theorem we obtain the existence of the conditional analytic Feynman integral for a class of functions that are not necessarily in $S(\nu)$.

THEOREM 3. *Let $\psi:\mathbb{R}^\nu \to \mathbb{C}$ be a Borel measurable function such that*

$$\int_{\mathbb{R}^\nu} |\psi(\vec{\eta})| \exp\{-a\|\vec{\eta}\|^2\}\, d\vec{\eta} < \infty$$

for all $a>0$. Let $F \in S(\nu)$ be given by (4). *Let $G(\vec{x}) = F(\vec{x})\psi(\vec{x}(t))$ and let $X(\vec{x}) = \vec{x}(t)$. Then $E^{\text{anf}_q}(G|X)$ exists for all real $q \neq 0$ and*

$$E^{\text{anf}_q}(G|X)(\vec{\eta}) = \psi(\vec{\eta}) E^{\text{anf}_q}(F|X)(\vec{\eta}) \tag{13}$$

for almost all $\vec{\eta} \in \mathbb{R}^\nu$.

Proof. We first note that

$$\int_{C^\nu[0,t]} |G(\lambda^{-1/2}\vec{x})|\, dm^\nu(\vec{x}) < \infty$$

for all $\lambda > 0$ since F is bounded. The existence of $E^{\text{anf}_q}(G|X)$ and (13) now follows from (3) (with F replaced by G) and Theorem 1.

Remark 4. Note that G is not necessarily in $S(\nu)$ since ψ may be unbounded.

4. The integral equation. Let $\mathcal{G}$ be the set of all $\mathbb{C}$-valued functions on $[0,\infty)\times\mathbb{R}^\nu$ of the form

$$\theta(s, \vec{U}) = \int_{\mathbb{R}^\nu} \exp\{i\langle \vec{U}, \vec{V}\rangle\}\, d\mu_s(\vec{V}) \tag{14}$$

where $\{\mu_S: 0 \leqq s < \infty\}$ is a family from $M(\mathbb{R}^\nu)$ satisfying the following two conditions:

(i) For every $B \in \mathcal{B}(\mathbb{R}^\nu)$, $\mu_S(B)$ is Borel measurable in s.

(ii) $\|\mu_S\| \in L_1[0, t]$ for all $t > 0$.

In [19], it is noted that $\theta(s, \vec{U})$ and $\theta(s, \vec{x}(s))$ are Borel measurable and that $F(\vec{x}) = \exp\{\int_0^t \theta(s, \vec{x}(s))\, ds\}$ is in $S(\nu)$.

Remark 5. Note that $S(\nu)$ actually depends on t and so we could write $S(\nu) = S_t(\nu)$. But for $0 < s \leqq t$, $F \in S_S(\nu) \Rightarrow F \in S_t(\nu)$. So we will usually delete the subscript unless the meaning is unclear without it.

THEOREM 4. *Let $\theta \in \mathcal{G}$ be given by* (14), *let*

$$F(\vec{x}) = F_t(\vec{x}) = \exp\left\{\int_0^t \theta(s, \vec{x}(s))\, ds\right\},$$

and let $X(\vec{x}) = X_t(\vec{x}) = \vec{x}(t)$. For $(t, \vec{\eta}, \lambda) \in (0,\infty)\times\mathbb{R}^\nu\times(0,\infty)$ let

$$H(t, \vec{\eta}, \lambda) = \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{-\frac{\lambda}{2t}\|\vec{\eta}\|^2\right\} E(F(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot))(\vec{\eta}). \tag{15}$$

Then for $(t, \vec{\eta}, \lambda) \in (0,\infty)\times\mathbb{R}^\nu\times(0,\infty)$, $H(t, \vec{\eta}, \lambda)$ satisfies the integral equation

$$\begin{aligned} H(t, \vec{\eta}, \lambda) &= \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{-\frac{\lambda}{2t}\|\vec{\eta}\|^2\right\} \\ &\quad + \int_0^t \left(\frac{\lambda}{2\pi(t-s)}\right)^{\nu/2} \int_{\mathbb{R}^\nu} \theta(s, \vec{\xi}) H(s, \vec{\xi}, \lambda) \\ &\quad \cdot \exp\left\{-\frac{\lambda}{2(t-s)}\|\vec{\eta}-\vec{\xi}\|^2\right\} d\vec{\xi}\, ds. \end{aligned} \tag{16}$$

Proof. Let $(t, \vec{\eta}, \lambda) \in (0, \infty) \times \mathbb{R}^\nu \times (0, \infty)$ be given. Then by differentiating the function $\exp\{\int_0^s \theta(u, \lambda^{-1/2}\vec{x}(u))\, du\}$ with respect to s and then integrating the derivative on $[0, t]$ we obtain the formula

$$\exp\left\{\int_0^t \theta(s, \lambda^{-1/2}\vec{x}(s))\, ds\right\} = 1 + \int_0^t \theta(s, \lambda^{-1/2}\vec{x}(s)) \exp\left\{\int_0^s \theta(u, \lambda^{-1/2}\vec{x}(u))\, du\right\} ds.$$

Hence, taking conditional expectations, using (3) and the Fubini theorem, we obtain

$$\begin{aligned}
&E(F(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot))(\vec{\eta}) \\
&\quad = 1 + E\left(\int_0^t \theta(s, \lambda^{-1/2}\vec{x}(s)) \exp\left\{\int_0^s \theta(u, \lambda^{-1/2}\vec{x}(u))\, du\right\} ds | \lambda^{-1/2}\vec{x}(t) = \vec{\eta}\right) \\
&\quad = 1 + \int_{C_0^\nu[0,t]} \int_0^t \theta\left(s, \lambda^{-1/2}\left[\vec{x}(s) - \frac{s}{t}\vec{x}(t)\right] + \frac{s}{t}\vec{\eta}\right) \\
&\qquad \cdot \exp\left\{\int_0^s \theta\left(u, \lambda^{-1/2}\left[\vec{x}(u) - \frac{u}{s}\vec{x}(s)\right]\right.\right. \\
&\qquad\qquad \left.\left. + \frac{u}{s}\left[\lambda^{-1/2}\left(\vec{x}(s) - \frac{s}{t}\vec{x}(t)\right) + \frac{s}{t}\vec{\eta}\right]\right) du\right\} ds\, dm^\nu(\vec{x}) \\
&\quad = 1 + \int_0^t \int_{C_0^\nu[0,t]} \theta\left(s, \lambda^{-1/2}\left[\vec{x}(s) - \frac{s}{t}\vec{x}(t)\right] + \frac{s}{t}\vec{\eta}\right) \\
&\qquad \cdot \exp\left\{\int_0^s \theta\left(u, \lambda^{-1/2}\left[\vec{x}(u) - \frac{u}{s}\vec{x}(s)\right]\right.\right. \\
&\qquad\qquad \left.\left. + \frac{u}{s}\left[\lambda^{-1/2}\left(\vec{x}(s) - \frac{s}{t}\vec{x}(t)\right) + \frac{s}{t}\vec{\eta}\right]\right) du\right\} dm^\nu(\vec{x})\, ds.
\end{aligned}$$

But $\lambda^{-1/2}[\vec{x}(s) - (s/t)\vec{x}(t)] + (s/t)\vec{\eta}$ is a Gaussian random variable with mean $(s/t)\vec{\eta}$ and variance $s(t-s)/\lambda t$ for $0 \leqq s < t$. Also it is independent of the random variable $\vec{x}(u) - (u/s)\vec{x}(s)$ for $0 < u < s < t$. But Brownian motion $\vec{x}(u)$ has stationary increments and so using (15) we obtain

$$\begin{aligned}
H(t, \vec{\eta}, \lambda) &= \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{-\frac{\lambda}{2t}\|\vec{\eta}\|^2\right\} \\
&\quad + \int_0^t \left(\frac{\lambda}{2\pi(t-s)}\right)^{\nu/2} \int_{\mathbb{R}^\nu} \theta(s, \vec{\xi}) \exp\left\{-\frac{\lambda}{2(t-s)}\|\vec{\eta} - \vec{\xi}\|^2\right\} \\
&\quad \cdot \left(\frac{\lambda}{2\pi s}\right)^{\nu/2} \exp\left\{-\frac{\lambda}{2s}\|\vec{\xi}\|^2\right\} E(F_s(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot))(\vec{\xi})\, d\vec{\xi}\, ds \\
&= \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{-\frac{\lambda}{2t}\|\vec{\eta}\|^2\right\} \\
&\quad + \int_0^t \left(\frac{\lambda}{2\pi(t-s)}\right)^{\nu/2} \int_{\mathbb{R}^\nu} \theta(s, \vec{\xi}) H(s, \xi, \lambda) \\
&\quad \cdot \exp\left\{-\frac{\lambda}{2(t-s)}\|\vec{\eta} - \vec{\xi}\|^2\right\} d\vec{\xi}\, ds,
\end{aligned}$$

which concludes the proof of Theorem 4.

THEOREM 5. *Let F and X be as in Theorem 4. Then for $(t, \vec{\eta}, \lambda) \in (0, \infty) \times \mathbb{R}^\nu \times \mathbb{C}^+$ the function*

$$H(t, \vec{\eta}, \lambda) = \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{-\frac{\lambda}{2t}\|\vec{\eta}\|^2\right\} E^{\mathrm{anw}_\lambda}(F|X)(\vec{\eta}) \tag{17}$$

satisfies the integral equation (16).

Proof. By Theorem 4 we know that $H(t, \vec{\eta}, \lambda)$ given by (17) satisfies the integral equation for $\lambda > 0$. Thus it suffices to show that both sides of (16) are analytic functions of λ throughout $\mathbb{C}^+$. But by Theorem 1, $H(t, \vec{\eta}, \lambda)$ exists and is analytic throughout $\mathbb{C}^+$. Thus it suffices to show that the second term on the right-hand side of (16), which we will denote by $h(\lambda)$, is an analytic function of λ on $\mathbb{C}^+$. We will use Morera's theorem to show that $h(\lambda)$ is analytic. First an application of the Dominated Convergence Theorem shows that $h(\lambda)$ is continuous on $\mathbb{C}^+$; an appropriate dominating function is obtained almost exactly as in the following argument and so will be omitted here. Now let Δ be a triangular path in $\mathbb{C}^+$. We need only show that $\int_\Delta h(\lambda)\, d\lambda = 0$. But this will clearly follow from the Cauchy Integral Theorem if we can justify moving the integral with respect to λ inside the other two integrals. Let $D = \sup\{|\lambda|: \lambda \in \Delta\}$ and $E = \inf\{\mathrm{Re}\,\lambda: \lambda \in \Delta\}$. Then

$$\left(\frac{D}{E}\right)^{\nu/2} \left(\frac{E}{2\pi(t-s)}\right)^{\nu/2} \|\mu_s\| \left(\frac{D}{2\pi t}\right)^{\nu/2} \exp\left\{-\frac{E}{2t}\|\vec{\eta}\|^2\right\} \|\sigma\| \exp\left\{-\frac{E}{2(t-s)}\|\vec{\eta} - \vec{\xi}\|^2\right\}$$

is a dominating function that is integrable with respect to $(s, \vec{\xi}, \lambda)$ on $[0, t] \times \mathbb{R}^\nu \times \Delta$.

THEOREM 6. *Let F and X be as in Theorem 4. Then for $(t, \vec{\eta}, q) \in (0, \infty) \times \mathbb{R}^\nu \times (\mathbb{R} - \{0\})$ the function*

$$H(t, \vec{\eta}, -qi) = \left(\frac{q}{2\pi i t}\right)^{\nu/2} \exp\left\{\frac{qi}{2t}\|\vec{\eta}\|^2\right\} E^{\mathrm{anf}_q}(F|X)(\vec{\eta}) \tag{18}$$

satisfies the Schrödinger integral equation

$$\begin{aligned} H(t, \vec{\eta}, -qi) &= \left(\frac{q}{2\pi i t}\right)^{\nu/2} \exp\left\{\frac{qi}{2t}\|\vec{\eta}\|^2\right\} \\ &\quad + \int_0^t \left(\frac{q}{2\pi i(t-s)}\right)^{\nu/2} \int_{\mathbb{R}^\nu} \theta(s, \vec{\xi}) H(s, \vec{\xi}, -qi) \\ &\quad \cdot \exp\left\{\frac{qi}{2(t-s)}\|\vec{\eta} - \vec{\xi}\|^2\right\} d\vec{\xi}\, ds. \end{aligned} \tag{19}$$

Proof. We first note that the techniques used in Theorem 5 will not work here since $\lim_{\lambda \to -qi} (|\lambda|/\mathrm{Re}\,\lambda) = +\infty$ and so it is not possible to find an integrable dominating function. Next note that by Theorem 1,

$$\lim_{\lambda \to -qi} H(t, \vec{\eta}, \lambda) = H(t, \vec{\eta}, -qi). \tag{20}$$

Next let

$$G(s, \vec{\xi}, \lambda) = \left(\frac{\lambda}{2\pi(t-s)}\right)^{\nu/2} \theta(s, \vec{\xi}) H(s, \vec{\xi}, \lambda) \exp\left\{-\frac{\lambda}{2(t-s)}\|\vec{\eta} - \vec{\xi}\|^2\right\}$$

for $s \in (0, t)$, $\vec{\xi} \in \mathbb{R}^\nu$ and $\lambda \neq 0$ such that $\mathrm{Re}\,\lambda \geqq 0$. Because the integral equation (16)

holds for all $\lambda \in \mathbb{C}^+$ it suffices in view of (20) to show that

$$\lim_{\lambda \to -qi} \int_0^t \int_{\mathbb{R}^\nu} G(s, \vec{\xi}, \lambda)\, d\vec{\xi}\, ds = \int_0^t \overline{\int_{\mathbb{R}^\nu}} G(s, \vec{\xi}, -qi)\, d\vec{\xi}\, ds. \tag{21}$$

But this follows from the following calculations provided that the use of the Dominated Convergence Theorem can be justified in steps 3, 4, and 6 below:

$$\begin{aligned}
\int_0^t \overline{\int_{\mathbb{R}^\nu}} G(s, \vec{\xi}, -iq)\, d\vec{\xi}\, ds &= \int_0^t \lim_{A\to\infty} \int_{\mathbb{R}^\nu} G(s, \vec{\xi}, -iq) \exp\left\{-\frac{\|\vec{\xi}\|^2}{2A}\right\} d\vec{\xi}\, ds \\
&= \int_0^t \lim_{A\to\infty} \int_{\mathbb{R}^\nu} \lim_{\lambda\to -iq} G(s, \vec{\xi}, \lambda) \exp\left\{-\frac{\|\vec{\xi}\|^2}{2A}\right\} d\vec{\xi}\, ds \\
&= \int_0^t \lim_{A\to\infty} \lim_{\lambda\to -iq} \int_{\mathbb{R}^\nu} G(s, \vec{\xi}, \lambda) \exp\left\{-\frac{\|\vec{\xi}\|^2}{2A}\right\} d\vec{\xi}\, ds \\
&= \int_0^t \lim_{\lambda\to -iq} \lim_{A\to\infty} \int_{\mathbb{R}^\nu} G(s, \vec{\xi}, \lambda) \exp\left\{-\frac{\|\vec{\xi}\|^2}{2A}\right\} d\vec{\xi}\, ds \\
&= \int_0^t \lim_{\lambda\to -iq} \int_{\mathbb{R}^\nu} G(s, \vec{\xi}, \lambda)\, d\vec{\xi}\, ds \\
&= \lim_{\lambda\to -iq} \int_0^t \int_{\mathbb{R}^\nu} G(s, \vec{\xi}, \lambda)\, d\vec{\xi}\, ds.
\end{aligned}$$

To show that we can apply the Dominated Convergence Theorem we will find a function that dominates the function

$$L(s) = \int_{\mathbb{R}^\nu} G(s, \vec{\xi}, \lambda) \exp\left\{-\frac{\|\vec{\xi}\|^2}{2A}\right\} d\vec{\xi}$$

for all A sufficiently large, for all $\lambda \in \mathbb{C}^+$ sufficiently close to $-qi$, and that is in $L_1[0, t]$ as a function of s. Actually we will find a dominating function that is independent of A and dominates for all $\lambda \neq 0$ such that $\operatorname{Re} \lambda \geqq 0$ and $|\lambda| \leqq \lambda_0 = 2|q| + 1$. Also it suffices to take the limit as $\lambda \to -qi$ along a horizontal line since we know that the limit in (21) exists.

First using the definitions of $L(s)$ and $G(s, \vec{\xi}, \lambda)$, and then (14) and (17), we obtain

$$\begin{aligned}
L(s) = \left(\frac{\lambda}{2\pi(t-s)}\right)^{\nu/2} \left(\frac{\lambda}{2\pi s}\right)^{\nu/2} \int_{\mathbb{R}^\nu} \left[\int_{\mathbb{R}^\nu} e^{i\langle \vec{\xi}, \vec{W}\rangle}\, d\mu_s(\vec{W})\right] \\
\cdot \exp\left\{-\frac{\lambda}{2(t-s)} \|\vec{\eta} - \vec{\xi}\|^2\right\} \exp\left\{-\frac{(s + A\lambda)}{2As} \|\vec{\xi}\|^2\right\} \\
\cdot \int_{L_2^\nu[0,s]} \exp\left\{-\frac{1}{2\lambda s} \sum_{j=1}^{\nu} [s\|v_j\|^2 - b_j^2] + \frac{i}{s}\langle \vec{\xi}, \vec{B}\rangle\right\} d\sigma_s(\vec{V})\, d\vec{\xi}
\end{aligned}$$

where $b_j = \int_0^s v_j(\tau)\, d\tau$, $\vec{B} = (b_1, \cdots, b_\nu)$ and $\sigma_s \in M(L_2^\nu[0, s])$ is such that

$$\exp\left\{\int_0^s \theta(\tau, \vec{x}(\tau))\, d\tau\right\} = \int_{L_2^\nu[0,s]} \exp\left\{i \sum_{j=1}^{\nu} \int_0^s v_j(\tau)\, d\tilde{x}_j(\tau)\right\} d\sigma_s(\vec{V}).$$

Next we use the Fubini theorem, then carry out the integration with respect to $\vec{\xi}$, simplify, and obtain

$$L(s)=\left(\frac{\lambda}{2\pi t}\right)^{\nu/2}\left(\frac{A\lambda t}{s(t-s)+A\lambda t}\right)^{\nu/2}\int_{\mathbb{R}^\nu}\int_{L_2^\nu[0,t]}\exp\left\{-\frac{1}{2\lambda s}\sum_{j=1}^{\nu}[s\|v_j\|^2-b_j^2]\right\}$$
$$\cdot\exp\left\{-\frac{\lambda}{2(t-s)}\|\vec{\eta}\|^2-\frac{As(t-s)}{2[s(t-s)+A\lambda t]}\left\|\vec{W}+\frac{\vec{B}}{s}-\frac{i\lambda\vec{\eta}}{t-s}\right\|^2\right\}d\sigma_s(\vec{V})\,d\mu_s(\vec{W}).$$

Now we claim that for all $A>0$ and all $\lambda\neq 0$ such that $\operatorname{Re}\lambda\geqq 0$ and $|\lambda|\leqq\lambda_0=2|q|+1$,

$$|L(s)|\leqq\left(\frac{\lambda_0}{2\pi t}\right)^{\nu/2}\|\sigma_s\|\,\|\mu_s\|\leqq\left(\frac{\lambda_0}{2\pi t}\right)^{\nu/2}\|\sigma\|\,\|\mu_s\|. \tag{22}$$

Once this claim is established, the proof is complete because the expression on the right-hand side of (22) is in $L_1[0,t]$ as a function of s.

(i) Clearly, $|A\lambda t/(s(t-s)+A\lambda t)|\leqq 1$ for all $A>0$ and $\operatorname{Re}\lambda\geqq 0$.

(ii) $|\exp\{-(1/2\lambda s)\sum_{j=1}^{\nu}[s\|v_j\|^2-b_j^2]\}|\leqq 1$ by the Cauchy-Schwarz inequality since $\operatorname{Re}(-1/2\lambda s)\leqq 0$.

(iii) Formula (22) will follow once we show that

$$\left|\exp\left\{-\frac{\lambda}{2(t-s)}\|\vec{\eta}\|^2-\frac{As(t-s)}{2[s(t-s)+A\lambda t]}\left\|\vec{W}+\frac{\vec{B}}{s}-\frac{i\lambda\vec{\eta}}{t-s}\right\|^2\right\}\right|\leqq 1$$

for all appropriate values of the variables involved. It suffices to consider one coordinate at a time and show that the real part of the exponent is nonpositive. We will work with the jth coordinate and to simplify notation we let $\eta_j=\eta$, $b_j=b$, and $w_j=w$. Then, recalling that $\lambda=p-qi$ and letting $y=w+b/s-q\eta/(t-s)$, we obtain

$$\operatorname{Re}\left\{-\frac{\lambda\eta^2}{2(t-s)}-\frac{As(t-s)}{2[s(t-s)+A\lambda t]}\left(w+\frac{b}{s}-\frac{i\lambda\eta}{t-s}\right)^2\right\}$$
$$=-\frac{As(t-s)[s(t-s)+Apt]}{2[(Apt+s(t-s))^2+(Aqt)^2]}\left[y+\frac{p\eta Aqt}{(t-s)[Apt+s(t-s)]}\right]^2$$
$$-\frac{p\eta^2}{2(t-s)}\left\{1-\frac{A^3spq^2t^2+Asp[Apt+s(t-s)]}{[(Apt+s(t-s))^2+(Atq)^2][Apt+s(t-s)]}\right\},$$

which is nonpositive since for $p\geqq 0$ and $0\leqq s\leqq t$,

$$A^3spq^2t^2+Asp[s(t-s)+Apt]^2\leqq[s(t-s)+Apt][(Apt+s(t-s))^2+(Atq)^2].$$

As a consequence of our next theorem we will see that $H(t,-\vec{\eta},-qi)$ is a fundamental solution to the Schrödinger partial differential equation.

THEOREM 7. *Let F and X be as in Theorem* 1. *Let $\psi\in\hat{M}(\mathbb{R}^\nu)$; that is to say*

$$\psi(\vec{\eta})=\int_{\mathbb{R}^\nu}\exp\{i\langle\vec{\eta},\vec{U}\rangle\}\,d\phi(\vec{U}) \tag{23}$$

for some $\phi\in M(\mathbb{R}^\nu)$. For $(t,\vec{\eta})\in(0,\infty)\times\mathbb{R}^\nu$, let

$$G(\vec{x})\equiv G_{t,\vec{\eta}}(\vec{x})=F(\vec{x})\psi(\vec{x}(t)+\vec{\eta}).$$

Then for all real $q\neq 0$ we have that

$$\Gamma(t,\vec{\eta},q)\equiv E^{\mathrm{anf}_q}(G)$$
$$=\int_{L_2^\nu[0,t]}\left[\exp\left\{-\frac{i}{2qt}\sum_{j=1}^{\nu}(t\|v_j\|^2-b_j^2)\right\}\right. \tag{24}$$

$$\cdot \int_{\mathbb{R}^\nu} \exp\left\{ i\langle \vec{\eta}, \vec{U}\rangle - \frac{i}{2qt}\|\vec{B}+t\vec{U}\|^2 \right\} d\phi(\vec{U}) \Big] d\sigma(\vec{V}).$$

In addition, we have the alternative expression

$$E^{\mathrm{anf}_q}(G) = \overline{\int_{\mathbb{R}^\nu}} E^{\mathrm{anf}_q}(F|X)(\vec{\xi}-\vec{\eta}) \left(\frac{q}{2\pi it}\right)^{\nu/2} \exp\left\{ \frac{iq\|\vec{\xi}-\vec{\eta}\|^2}{2t} \right\} \psi(\vec{\xi})\, d\vec{\xi} \tag{25}$$

where $E^{\mathrm{anf}_q}(F|X)(\cdot)$ *is given by* (7).

Proof. Since $P_{X(\cdot)+\vec{\eta}}(d\vec{\xi}) = (2\pi t)^{-\nu/2} \exp\{-\|\vec{\xi}-\vec{\eta}\|^2/2t\}\, d\vec{\xi}$, by Lemma 1 of [24] it follows that

$$\begin{aligned} J(\lambda) &= \int_{C_0^\nu[0,t]} G(\lambda^{-1/2}\vec{x})\, dm^\nu(\vec{x}) \\ &= \int_{C_0^\nu[0,t]} F(\lambda^{-1/2}\vec{x})\psi(X(\lambda^{-1/2}\vec{x})+\vec{\eta})\, dm^\nu(\vec{x}) \\ &= \int_{\mathbb{R}^\nu} E(F(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot)+\vec{\eta})(\vec{\xi})\psi(\vec{\xi}) \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{ -\frac{\lambda\|\vec{\xi}-\vec{\eta}\|^2}{2t} \right\} d\vec{\xi} \\ &= \int_{\mathbb{R}^\nu} E(F(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot))(\vec{\xi}-\vec{\eta})\psi(\vec{\xi}) \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{ -\frac{\lambda\|\vec{\xi}-\vec{\eta}\|^2}{2t} \right\} d\vec{\xi} \end{aligned}$$

for all $\lambda > 0$. Then, using Theorem 1 and Morera's theorem, we obtain

$$E^{\mathrm{anw}_\lambda}(G) = \int_{\mathbb{R}^\nu} E^{\mathrm{anw}_\lambda}(F|X)(\vec{\xi}-\vec{\eta}) \left(\frac{\lambda}{2\pi t}\right)^{\nu/2} \exp\left\{ -\frac{\lambda\|\vec{\xi}-\vec{\eta}\|^2}{2t} \right\} \psi(\vec{\xi})\, d\vec{\xi} \tag{26}$$

for all $\lambda \in \mathbb{C}^+$. Next we substitute for $E^{\mathrm{anw}_\lambda}(F|X)(\vec{\xi}-\vec{\eta})$ and $\psi(\vec{\xi})$ in (26) using (7) and (23), use the Fubini theorem, and then carry out the integration with respect to $\vec{\xi}$ and obtain the formula

$$\begin{aligned} E^{\mathrm{anw}_\lambda}(G) = \int_{L_2^\nu[0,t]} \Bigg[&\exp\left\{ -\frac{1}{2\lambda t} \sum_{j=1}^{\nu} (t\|v_j\|^2 - b_j^2) \right\} \\ &\cdot \int_{\mathbb{R}^\nu} \exp\left\{ i\langle \vec{\eta}, \vec{U}\rangle - \frac{1}{2\lambda t}\|\vec{B}+t\vec{U}\|^2 \right\} d\phi(\vec{U}) \Bigg] d\sigma(\vec{V}) \end{aligned} \tag{27}$$

for all $\lambda \in \mathbb{C}^+$. Next we note that the right-hand side of (27) is continuous in λ for $\mathrm{Re}\,\lambda \geqq 0$, $\lambda \neq 0$, and hence $E^{\mathrm{anf}_q}(G)$ exists and is given by (24).

To obtain the alternative expression (25), we use (10), (7), the Dominated Convergence Theorem, and (24):

$$\begin{aligned} &\overline{\int_{\mathbb{R}^\nu}} E^{\mathrm{anf}_q}(F|X)(\vec{\xi}-\vec{\eta}) \left(\frac{q}{2\pi it}\right)^{\nu/2} \exp\left\{ \frac{iq}{2t}\|\vec{\xi}-\vec{\eta}\|^2 \right\} \psi(\vec{\xi})\, d\vec{\xi} \\ &= \lim_{A\to+\infty} \int_{\mathbb{R}^\nu} E^{\mathrm{anf}_q}(F|X)(\vec{\xi}-\vec{\eta}) \left(\frac{q}{2\pi it}\right)^{\nu/2} \exp\left\{ \frac{iq}{2t}\|\vec{\xi}-\vec{\eta}\|^2 - \frac{1}{2A}\|\vec{\xi}\|^2 \right\} \psi(\vec{\xi})\, d\vec{\xi} \\ &= \lim_{A\to+\infty} \int_{\mathbb{R}^\nu} \Bigg[\int_{L_2^\nu[0,t]} \exp\left\{ -\frac{i}{2qt} \sum_{j=1}^{\nu} (t\|v_j\|^2 - b_j^2) + \frac{i}{t}\langle \vec{\xi}-\vec{\eta}, \vec{B}\rangle \right\} d\sigma(\vec{V}) \\ &\qquad \cdot \left(\frac{q}{2\pi it}\right)^{\nu/2} \exp\left\{ \frac{iq}{2t}\|\vec{\xi}-\vec{\eta}\|^2 - \frac{1}{2A}\|\vec{\xi}\|^2 \right\} \int_{\mathbb{R}^\nu} \exp\{i\langle \vec{\xi}, \vec{U}\rangle\}\, d\phi(\vec{U}) \Bigg] d\vec{\xi} \end{aligned}$$

$$= \lim_{A\to+\infty} \int_{L_2^\nu[0,t]} \left[\int_{\mathbb{R}^\nu} \exp\left\{ -\frac{i}{2qt} \sum_{j=1}^{\nu} (t\|v_j\|^2 - b_j^2) - \frac{i}{t}\langle \vec{\eta}, \vec{B}\rangle \right\} \left(\frac{q}{2\pi i t}\right)^{\nu/2} \right.$$

$$\left. \cdot \left[\int_{\mathbb{R}^\nu} \exp\left\{ \frac{i}{t}\langle \vec{\xi}, \vec{B}\rangle + \frac{iq}{2t}\|\vec{\xi} - \vec{\eta}\|^2 - \frac{1}{2A}\|\vec{\xi}\|^2 + i\langle \vec{\xi}, \vec{U}\rangle \right\} d\vec{\xi} \right] d\phi(\vec{U}) \right] d\sigma(\vec{V})$$

$$= \lim_{A\to+\infty} \int_{L_2^\nu[0,t]} \left[\int_{\mathbb{R}^\nu} \exp\left\{ -\frac{i}{2qt} \sum_{j=1}^{\nu} (t\|v_j\|^2 - b_j^2) - \frac{i}{t}\langle \vec{\eta}, \vec{B}\rangle + \frac{iq}{2t}\|\vec{\eta}\|^2 \right\} \left(\frac{q}{2\pi i t}\right)^{\nu/2} \right.$$

$$\left. \cdot \left[\int_{\mathbb{R}^\nu} \exp\left\{ -\frac{(t - Aiq)}{2At}\|\vec{\xi}\|^2 + \frac{i}{t}\langle \vec{\xi}, \vec{B} - q\vec{\eta} + t\vec{U}\rangle \right\} d\vec{\xi} \right] d\phi(\vec{U}) \right] d\sigma(\vec{V})$$

$$= \lim_{A\to+\infty} \int_{L_2^\nu[0,t]} \left[\exp\left\{ -\frac{i}{2qt} \sum_{j=1}^{\nu} (t\|v_j\|^2 - b_j^2) - \frac{i}{t}\langle \vec{\eta}, \vec{B}\rangle + \frac{iq}{2t}\|\vec{\eta}\|^2 \right\} \right.$$

$$\left. \cdot \left(\frac{q}{2\pi i t}\right)^{\nu/2} \left(\frac{2\pi At}{t - Aiq}\right)^{\nu/2} \int_{\mathbb{R}^\nu} \exp\left\{ -\frac{A}{2t(t - Aiq)}\|\vec{B} - q\vec{\eta} + t\vec{U}\|^2 \right\} d\phi(\vec{U}) \right] d\sigma(\vec{V})$$

$$= \int_{L_2^\nu[0,t]} \left[\exp\left\{ -\frac{i}{2qt} \sum_{j=1}^{\nu} (t\|v_j\|^2 - b_j^2) \right\} \right.$$

$$\left. \cdot \int_{\mathbb{R}^\nu} \exp\left\{ i\langle \vec{\eta}, \vec{U}\rangle - \frac{i}{2qt}\|\vec{B} + t\vec{U}\|^2 \right\} d\phi(\vec{U}) \right] d\sigma(\vec{V})$$

$$= E^{\mathrm{anf}_q}(G).$$

Remark 6. Note that (25) can be written as $\Gamma(t, \vec{\eta}, q) = H(t, -(\cdot), -qi) * \psi(\vec{\eta})$ where * denotes convolution and $H(t, \vec{\eta}, -qi)$ is given by (18). Next consider the Schrödinger partial differential equation:

$$i\hbar \frac{\partial \Gamma}{\partial t} = -\frac{\hbar^2}{2m}\Delta\Gamma + \theta(t, \vec{\eta})\Gamma, \qquad \Gamma(0, \vec{\eta}) = \psi(\vec{\eta}) \tag{28}$$

where Δ is the Laplacian on $\mathbb{R}^\nu$, $\hbar = h/2\pi$ where h is Planck's constant and $\theta(s, \cdot)$ is a time-dependent potential.

If $\theta(s, \cdot) = \theta(\cdot)$ is a time-independent and an $\mathbb{R}$-valued potential in $\hat{M}(\mathbb{R}^\nu)$, and ψ is in $\hat{M}(\mathbb{R}^\nu) \cap L_2(\mathbb{R}^\nu)$, then by the results of [15], [20] combined with the results of [1, Thm. 3.1], it follows that $E^{\mathrm{anf}_q}(F(\vec{x})\psi(\vec{x}(t) + \vec{\eta}))$ (with $q = m/\hbar$) for some F in $S(\nu)$ is a (weak) solution of the Schrödinger equation (28). Hence, (25) shows that $H(t, -\vec{\eta}, -qi)$ is a fundamental solution to the Schrödinger equation (28).

If $\theta(s, \vec{\eta})$ is the time-dependent potential given by (14), then by [19, Thm. 7.1] $\Gamma(t, \vec{\eta}, q)$ is a solution of the following integral equation that is formally equivalent to the Schrödinger equation (28) (with $q = m/\hbar$):

$$\Gamma(t, \vec{\eta}, q) = \left(\frac{q}{2\pi i t}\right)^{\nu/2} \int_{\mathbb{R}^\nu} \psi(\vec{U}) \exp\left\{ \frac{iq\|\vec{\eta} - \vec{U}\|^2}{2t} \right\} d\vec{U}$$

$$+ \int_0^t \left[\frac{q}{2\pi i(t - s)} \right]^{\nu/2} \int_{\mathbb{R}^\nu} \theta(s, \vec{U})\Gamma(s, \vec{U}, q) \exp\left\{ \frac{iq\|\vec{\eta} - \vec{U}\|^2}{2(t - s)} \right\} d\vec{U}\, ds.$$

Theorem 6 shows that the solution $\Gamma(t, \vec{\eta}, q)$ of this integral equation can be obtained by use of (25).

5. More general conditioning functions. In this section with $\nu = 1$ we obtain results corresponding to those in § 3 but for conditioning functions of the form $X(x) = (x(t_1), \cdots, x(t_n))$.

Let n be a positive integer and let $0 = t_0 < t_1 < \cdots < t_n = t$ be a partition of $[0, t]$. For each $x \in C_0[0, t]$, define the polygonal function $[x]$ on $[0, t]$ by

$$[x](s) = x(t_{k-1}) + \frac{s - t_{k-1}}{t_k - t_{k-1}}(x(t_k) - x(t_{k-1})), \quad t_{k-1} \leqq s \leqq t_k, \quad k = 1, \cdots, n.$$

Similarly, for each $\vec{\eta} = (\eta_1, \cdots, \eta_n) \in \mathbb{R}^n$, define the polygonal function $[\vec{\eta}]$ of $\vec{\eta}$ on $[0, t]$ by

$$[\vec{\eta}](s) = \eta_{k-1} + \frac{s - t_{k-1}}{t_k - t_{k-1}}(\eta_k - \eta_{k-1}), \quad t_{k-1} \leqq s \leqq t_k, \quad k = 1, \cdots, n.$$

As noted in [22], $\{x(s) - [x](s), t_{k-1} \leqq s \leqq t_k\}$, $k = 1, \cdots, n$ are independent Brownian bridge processes. Furthermore, the processes $\{x(s) - [x](s), 0 \leqq s \leqq t\}$ and $X(x) = (x(t_1), \cdots, x(t_n))$ are stochastically independent.

In [22] Park and Skoug have shown that if F is Borel measurable and Wiener integrable then the conditional Wiener integral $E(F(x)|X(x) \equiv (x(t_1), \cdots, x(t_n)))(\vec{\eta})$ can be expressed in terms of an ordinary Wiener integral by the formula

$$(29) \quad E(F(x)|X(x)) \equiv (x(t_1), \cdots, x(t_n))(\vec{\eta}) = \int_{C_0[0,t]} F(x - [x] + [\vec{\eta}])\, dm(x).$$

DEFINITION 3. Let $0 = t_0 < t_1 < \cdots < t_n = t$ be a partition of $[0, t]$. Then for each function $v \in L_2[0, t]$ we define the sectional average of v by letting

$$\bar{v}(s) = \frac{1}{t_k - t_{k-1}} \int_{t_{k-1}}^{t_k} v(u)\, du$$

on each subinterval $(t_{k-1}, t_k]$ and by letting $\bar{v}(0) = 0$.

Note that $\bar{v}$ is a step function of bounded variation on $[0, t]$. The following theorem gives a relationship involving $\bar{v}$ and $[x]$ that is very useful in computing conditional expectations.

THEOREM 8. *Let $v \in L_2[0, t]$. Then*

$$(30) \quad \int_0^t v(s)\bar{v}(s)\, ds = \int_0^t \bar{v}^2(s)\, ds,$$

$$(31) \quad \|v - \bar{v}\|_2^2 = \|v\|_2^2 - \|\bar{v}\|_2^2 \geqq 0, \quad \textit{and}$$

$$(32) \quad \int_0^t v(s)\, d[x](s) = \int_0^t \bar{v}(s)\, dx(s) = \int_0^t \bar{v}(s)\, d[x](s) \quad \textit{for each } x \in C_0[0, t].$$

Proof. Equation (30) follows easily from the definition of $\bar{v}$ while (31) follows from (30). To obtain (32) note that for each $k = 1, \cdots, n$,

$$\int_{t_{k-1}}^{t_k} v(s)\, d[x](s) = \frac{x(t_k) - x(t_{k-1})}{t_k - t_{k-1}} \int_{t_{k-1}}^{t_k} v(s)\, ds$$

$$= \bar{v}(t_k)(x(t_k) - x(t_{k-1})) = \int_{t_{k-1}}^{t_k} \bar{v}(s)\, dx(s).$$

THEOREM 9. *Let* $F(x)=\int_{L_2[0,t]}\exp\{i\int_0^t v(s)\,\tilde{d}x(s)\}\,d\sigma(v)$ *be an element of* $S(1)$ *and let* $X(x)=(x(t_1),\cdots,x(t_n))$. *Then for all* $\lambda\in\mathbb{C}^+$ *and each* $\vec{\eta}=(\eta_1,\cdots,\eta_n)\in\mathbb{R}^n$

$$E^{\mathrm{anw}_\lambda}(F|X)(\vec{\eta})=\int_{L_2[0,t]}\exp\left\{i\int_0^t v(s)\,d[\vec{\eta}](s)-\frac{1}{2\lambda}\int_0^t(v(s)-\bar{v}(s))^2\,ds\right\}d\sigma(v), \tag{33}$$

and for all real $q\neq 0$,

$$E^{\mathrm{anf}_q}(F|X)(\vec{\eta})=\int_{L_2[0,t]}\exp\left\{i\int_0^t v(s)\,d[\vec{\eta}](s)-\frac{i}{2q}\int_0^t(v(s)-\bar{v}(s))^2\,ds\right\}d\sigma(v). \tag{34}$$

Proof. Using (29), the Fubini theorem, and Theorem 8, we see that for each $\lambda>0$ and each $\vec{\eta}\in\mathbb{R}^n$,

$$\begin{aligned}
&E(F(\lambda^{-1/2}\cdot)|X(\lambda^{-1/2}\cdot))(\vec{\eta})\\
&\quad=\int_{C_0[0,t]}\int_{L_2[0,t]}\exp\left\{i\int_0^t v(s)\,\tilde{d}(\lambda^{-1/2}x(s)-\lambda^{-1/2}[x](s)\right.\\
&\qquad\qquad\left.+[\vec{\eta}](s))\right\}d\sigma(v)\,dm(x)\\
&\quad=\int_{L_2[0,t]}\exp\left\{i\int_0^t v(s)\,d[\vec{\eta}](s)\right\}\int_{C_0[0,t]}\\
&\qquad\cdot\exp\left\{i\lambda^{-1/2}\int_0^t(v(s)-\bar{v}(s))\,\tilde{d}x(s)\right\}dm(x)\,d\sigma(v) \qquad (35)\\
&\quad=\int_{L_2[0,t]}\exp\left\{i\int_0^t v(s)\,d[\vec{\eta}](s)\right\}\\
&\qquad\cdot(2\pi)^{-1/2}\int_{\mathbb{R}}\exp\{iu\lambda^{-1/2}\|v-\bar{v}\|-u^2/2\}\,du\,d\sigma(v)\\
&\quad=\int_{L_2[0,t]}\exp\left\{i\int_0^t v(s)\,d[\vec{\eta}](s)-\frac{1}{2\lambda}\int_0^t(v(s)-\bar{v}(s))^2\,ds\right\}d\sigma(v).
\end{aligned}$$

But since $\sigma\in M(L_2[0,t])$, it is not hard to see that the right-hand side of (35) above is an analytic function of λ throughout $\mathbb{C}^+$ and is a continuous function of λ for $\mathrm{Re}\,\lambda\geqq 0$, $\lambda\neq 0$.

THEOREM 10. *Let* F *and* X *be as in Theorem* 9. *Then for all* $\lambda\in\mathbb{C}^+$,

$$\begin{aligned}
&\int_{\mathbb{R}^n}\prod_{k=1}^n\left(\frac{\lambda}{2\pi(t_k-t_{k-1})}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}\sum_{k=1}^n\frac{(\eta_k-\eta_{k-1})^2}{t_k-t_{k-1}}\right\}E^{\mathrm{anw}_\lambda}(F|X)(\vec{\eta})\,d\vec{\eta}\\
&\quad=\int_{L_2[0,t]}\exp\left\{-\frac{1}{2\lambda}\|v\|^2\right\}d\sigma(v)=E^{\mathrm{anw}_\lambda}(F)
\end{aligned} \tag{36}$$

and for all real $q\neq 0$,

$$\begin{aligned}
&\int_{\mathbb{R}^n}\prod_{k=1}^n\left(\frac{q}{2\pi i(t_k-t_{k-1})}\right)^{1/2}\exp\left\{\frac{iq}{2}\sum_{k=1}^n\frac{(\eta_k-\eta_{k-1})^2}{t_k-t_{k-1}}\right\}E^{\mathrm{anf}_q}(F|X)(\vec{\eta})\,d\vec{\eta}\\
&\quad=\int_{L_2[0,t]}\exp\left\{-\frac{i}{2q}\|v\|^2\right\}d\sigma(v)=E^{\mathrm{anf}_q}(F)
\end{aligned} \tag{37}$$

where $E^{\mathrm{anw}_\lambda}(F|X)$ *and* $E^{\mathrm{anf}_q}(F|X)$ *are given by* (33) *and* (34), *respectively.*

Proof. Let $\lambda\in\mathbb{C}^+$ be given. Note that in view of (32) we have that

$$\int_0^t v(s)\,d[\vec{\eta}](s)=\int_0^t\bar{v}(s)\,d[\vec{\eta}](s)=\sum_{k=1}^n\frac{(\eta_k-\eta_{k-1})}{t_k-t_{k-1}}\int_{t_{k-1}}^{t_k}v(s)\,ds.$$

Thus using (33), the Fubini theorem, and Theorem 8, we see that the left-hand side of (36) equals the expression

$$\int_{L_2[0,t]} \exp\left\{-\frac{1}{2\lambda}\int_0^t (v(s)-\bar{v}(s))^2\,ds\right\} \prod_{k=1}^{n}\left(\frac{\lambda}{2\pi(t_k-t_{k-1})}\right)^{1/2}$$

$$\cdot \int_{\mathbb{R}^n} \exp\left\{-\frac{\lambda}{2}\sum_{k=1}^{n}\frac{(\eta_k-\eta_{k-1})^2}{(t_k-t_{k-1})}+i\sum_{k=1}^{n}\frac{(\eta_k-\eta_{k-1})}{(t_k-t_{k-1})}\int_{t_{k-1}}^{t_k} v(s)\,ds\right\} d\vec{\eta}\,d\sigma(v)$$

$$= \int_{L_2[0,t]} \exp\left\{-\frac{1}{2\lambda}\int_0^t (v(s)-\bar{v}(s))^2\,ds\right\}\left(\frac{\lambda}{2\pi}\right)^{n/2}$$

$$\cdot \int_{\mathbb{R}^n} \exp\left\{-\frac{\lambda}{2}\sum_{k=1}^{n}\xi_k+i\sum_{k=1}^{n}\xi_k(t_k-t_{k-1})^{-1/2}\int_{t_{k-1}}^{t_k} v(s)\,ds\right\} d\xi_1\cdots d\xi_n\,d\sigma(v)$$

$$= \int_{L_2[0,t]} \exp\left\{-\frac{1}{2\lambda}\int_0^t (v(s)-\bar{v}(s))^2\,ds\right.$$

$$\left.-\frac{1}{2\lambda}\sum_{k=1}^{n}\frac{1}{t_k-t_{k-1}}\left(\int_{t_{k-1}}^{t_k} v(s)\,ds\right)^2\right\} d\sigma(v)$$

$$= \int_{L_2[0,t]} \exp\left\{-\frac{1}{2}\|v\|_2^2+\frac{1}{2\lambda}\|\bar{v}\|_2^2-\frac{1}{2\lambda}\|\bar{v}\|_2^2\right\} d\sigma(v)$$

$$= \int_{L_2[0,t]} \exp\left\{-\frac{1}{2\lambda}\|v\|_2^2\right\} d\sigma(v),$$

which establishes (36). To establish (37), use the summation procedure (10) and proceed as above; carrying out the integration with respect to $\vec{\eta}=(\eta_1,\cdots,\eta_n)$ is rather long and tedious.

Remark 7. Theorems 9 and 10 have ν-dimensional counterparts. For example, if $F\in S(\nu)$ and if $X(\vec{x})=(\vec{x}(t_1),\cdots,\vec{x}(t_n))$ for $\vec{x}$ in $C_0^{\nu}[0,t]$, then the equation corresponding to (33) is

$$E(F(\vec{x})|X(\vec{x}))(\vec{\eta}_1,\cdots,\vec{\eta}_\nu)$$

$$= \int_{L_2^{\nu}[0,t]} \exp\left\{i\sum_{j=1}^{\nu}\int_0^t v_j(s)\,d[\vec{\eta}_j](s)-\frac{i}{2q}\sum_{j=1}^{\nu}\int_0^t (v_j(s)-\bar{v}_j(s))^2\,ds\right\} d\sigma(\vec{V})$$

where $\vec{V}=(v_1,\cdots,v_\nu)$ and $\vec{\eta}_j=(\eta_{j,1},\cdots,\eta_{j,n})$ for $j=1,\cdots,\nu$.

Acknowledgment. The first author wishes to express his gratitude to Professor G. W. Johnson for his encouragement and valuable advice as well as the hospitality of the University of Nebraska.

REFERENCES

[1] S. ALBEVERIO AND R. HØEGH-KROHN, *Mathematical Theory of Feynman Path Integrals*, Lecture Notes in Mathematics 523, Springer-Verlag, Berlin, New York, 1976.

[2] R. H. CAMERON AND D. A. STORVICK, *An operator valued function space integral and a related integral equation*, J. Math. Mech., 18 (1968), pp. 517-552.

[3] ———, *Some Banach algebras of analytic Feynman integrable functionals*, in Analytic Functions, Kozubnik, 1979, Lecture Notes in Mathematics 798, Springer-Verlag, Berlin, New York, 1980, pp. 18-67.

[4] ———, *Analytic Feynman integral solutions of an integral equation related to the Schroedinger equation*, J. Analyse Math., 38 (1980), pp. 34-66.

[5] R. H. CAMERON AND D. A. STORVICK, *A new translation theorem for the analytic Feynman integral*, Rev. Roumaine Math. Pures Appl., 27 (1982), pp. 937-944.
[6] ———, *A simple definition of the Feynman integral, with applications*, Mem. Amer. Math. Soc. 288, 46 (1983), pp. 1-46.
[7] ———, *New existence theorems and evaluation formulas for sequential Feynman integrals*, Proc. London Math. Soc., 52 (1986), pp. 557-581.
[8] K. S. CHANG AND J. S. CHANG, *Evaluation of some conditional Wiener integrals*, Bulletin Korean Math. Soc., 21 (1984), pp. 99-106.
[9] K. S. CHANG, G. W. JOHNSON, AND D. L. SKOUG, *The Feynman integral of quadratic potentials depending on two time variables*, Pacific J. Math., 122 (1986), pp. 11-33.
[10] ———, *Functions in the Fresnel class*, Proc. Amer. Math. Soc., 100 (1987), pp. 309-318.
[11] ———, *Necessary and sufficient conditions for membership in the Banach algebra S for certain classes of functions*, Supplemento ai Rendiconti del Circolo Matematico di Palermo, 17 (1987), pp. 153-171.
[12] D. M. CHUNG, *Scale-invariant measurability in abstract Wiener spaces*, Pacific J. Math., 130 (1987), pp. 27-40.
[13] D. M. CHUNG AND S. J. KANG, *Conditional Wiener integrals and an integral equation*, J. Korean Math. Soc., 25 (1988), pp. 37-52.
[14] D. ELWORTHY AND A. TRUMAN, *Feynman maps, Cameron-Martin formulae and anharmonic oscillators*, Ann. Inst. H. Poincaré, 41 (1984), pp. 115-142.
[15] G. W. JOHNSON, *The equivalence of two approaches to the Feynman integral*, J. Math. Phys., 23 (1982), pp. 2090-2096.
[16] G. W. JOHNSON AND D. L. SKOUG, *Scale-invariant measurability in Wiener space*, Pacific J. Math., 83 (1979), pp. 157-176.
[17] ———, *Notes on the Feynman integral*, I, Pacific J. Math., 93 (1981), pp. 313-324.
[18] ———, *Notes on the Feynman integral*, II, J. Funct. Anal., 41 (1981), pp. 277-289.
[19] ———, *Notes on the Feynman integral*, III: *the Schroedinger equation*, Pacific J. Math., 105 (1983), pp. 321-358.
[20] G. KALLIANPUR, D. KANNAN, AND R. L. KARANDIKAR, *Analytic and sequential Feynman integrals on abstract Wiener and Hilbert spaces and a Cameron-Martin formula*, Ann. Inst. H. Poincaré, 21 (1985), pp. 323-361.
[21] C. PARK AND D. L. SKOUG, *The Feynman integral of quadratic potentials depending on n time parameters*, Nagoya Math. J., 110 (1988), pp. 151-162.
[22] ———, *A simple formula for conditional Wiener integrals with applications*, Pacific J. Math., 135 (1988), pp. 381-394.
[23] J. YEH, *Inversion of conditional expectations*, Pacific J. Math., 52 (1974), pp. 631-640.
[24] ———, *Inversion of conditional Wiener integrals*, Pacific J. Math., 59 (1975), pp. 623-638.
[25] ———, *Transformation of conditional Wiener integrals under translation and the Cameron-Martin translation theorem*, Tôhoku Math. J. (2), 30 (1978), pp. 505-515.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 966–986, July 1989

012

THE RIEMANN–HILBERT PROBLEM AND INVERSE SCATTERING

XIN ZHOU†

Abstract. The connection between the Riemann–Hilbert factorization on self-intersecting contours and a class of singular integral equations is studied with a pair of decomposing algebras. This provides an effective way of treating the inverse scattering problem for first-order systems. We also show that the matrix functions with positive definite real parts on the real axis and Schwarz reflection invariant elsewhere only have zero partial indices. In particular, this implies the solvability for the inverse scattering problem with skew Schwarz reflection invariant system coefficients $J(z)$ and $q(\cdot, z)$. This includes, for instance, the system associated with the generalized sine-Gordon equation.

Key words. factorization, inverse scattering, partial indices

AMS(MOS) subject classifications. 34B25, 35Q15

1. Introduction.

Recent progress in the theory of the inverse scattering method motivates the study of the Riemann–Hilbert problem with self-intersecting contours. This paper shows that although it is inadequate to study the Riemann–Hilbert problem with this type of contours for a single decomposing algebra, we may still acquire, by working on a pair of decomposing algebras (see §9), the classical results which have been obtained in the case of nonself-intersecting contours [7]. This provides a new approach for the study of the inversion of the Beals–Coifman scattering data.

It is by now well understood [3], [6], [12], that the inverse scattering problem for the first order $n \times n$ system

$$\frac{d}{dx}\mathbf{m} - \operatorname{ad} J(z)\mathbf{m} = q(x, z)\mathbf{m} \tag{1.1}$$

for certain rational $J(z)$ and $q(\cdot, z)$ can be formulated as a Riemann–Hilbert problem with zero partial indices and a parameter x. The inverse scattering of AKNS system (2×2 and $J(z) = zJ$, $q(x, z) = q(x)$ [2]) has been formulated as a Gelfand–Levitan–Marchenko integral equation with its integral operator being compact and vanishing in norm as $x \to -\infty$ (or $+\infty$). These nice properties give the existence and the desired decay for the solution of the Gelfand–Levitan–Marchenko equation near $x = -\infty$ (or$+\infty$). They also aid in showing the solvability of the inverse scattering problem for skew Hermitian J and $q(x)$ [2], and the generic solvability for general J and $q(x)$. The solution of a Riemann–Hilbert problem with zero partial indices may also be obtained by solving a singular integral equation. However, the singular integral operator thereof need not have the nice properties stated above for the Gelfand–Levitan–Marchenko equation. Yet the complexity of the contours involved in the inverse scattering problem forbids the direct application of the existing general theory of the Riemann–Hilbert problem. With all of these difficulties, the inverse scattering problem for the $n \times n$ AKNS system was rigorously treated in [3] by splitting the problem, via a careful rational approximation, into a small norm problem of a singular integral equation and

*Received by the editors January 19, 1988; accepted for publication (in revised form) July 27, 1988.

†Department of Mathematics, University of Wisconsin–Madision, Madison, Wisconsin 53706. This work was done when the author was at the University of Rochester.

a linear algebraic problem. Thereafter the same method was applied to the inverse scattering for two other systems [1], [9].

In this paper we study the inverse scattering problem of system (1.1) for a class of $J(z)$ and $q(\cdot, z)$ (see §8) with a more direct method based on the Fredholm theory of singular integral operators. Our extended theory of the Riemann–Hilbert problem for decomposing algebras makes it possible to establish a singular integral equation which may be viewed as a natural extension of the Gelfand–Levitan–Marchenko equation. More precisely, we obtain a singular integral equation

$$m = I + C_x m \tag{1.2}$$

for all x real (also for all t real in the time evolution problem) with the following properties:

(1) The operator $Id - C_x$ admits an explicit regularization in the form

$$(Id - \tilde{C}_x)(Id - C_x) = Id + T_x \tag{1.3}$$

where $\tilde{C}_x$ is the operator for an associated Riemann–Hilbert problem and T_x is a compact operator.

(2) (The Fredholm alternative.) The Fredholm index of $Id - C_x$ is zero.

(3) $T_x \to 0$ in norm as $x \to -\infty$.

(4) When the scattering data (less I) is supported away from an open set containing the poles of $J(z)$, C_x is entire in x (also in t in the time evolution problem). We remark that for the 2×2 AKNS system the regularization is not necessary. In this case, on a Hardy space C_x in (1.2) is compact and vanishes in norm as $x \to -\infty$. In fact, the Fourier transform of (1.2) on the Hardy space is the Gelfand–Levitan–Marchenko equation.

Many useful results for the inverse scattering problem may be deduced directly from (1.2) with the above properties. For instance, properties 2 and 3 provide the invertibility of $Id - C_x$ near $x = -\infty$. The decay of $m - I$ near $x = -\infty$ depends on the oscillatory integral $C_x I$ by

$$m - I = (Id - C_x)^{-1} C_x I = O(C_x I).$$

Moreover, in case the scattering data is supported away from an open set containing the poles of $J(z)$, the analytic Fredholm theorem helps to invert $Id - C_x$. This enables us to give a proof of the generic solvability (Theorem 6.3) with an explicit perturbation. We point out here that the x-dependent data is not bounded as $x \to \infty$ in the Sobolev spaces. In [3] the x-dependent norms on Sobolev spaces were introduced to obtain the small norm problem on these spaces. By our method, this can be avoided by using the Fredholm alternative, which shows that if $Id - C_x$ is invertible on $\mathbf{L}^2$, then it is invertible on every space set theoretically contained in the span of the constant functions and $\mathbf{L}^2$ as long as the regularization is valid for such a space.

In addition, our method exhibits the following advantages in comparison with that introduced in [3]:

(1) The rational approximation used in [3] is local to the parameter t although for x it can be done once for $x \leq 0$ and once for $x \geq 0$. The produced algebraic systems for $x \leq 0$ and for $x \geq 0$ are not well connected. In case of Schwartz class scattering data (the scattering data and their arbitrary order derivatives having arbitrary orders of polynomial decay at all the poles of $J(z)$), to derive the arbitrary polynomial decay for corresponding potentials, infinitely many different algebraic systems are required. While (1.2) stands for x and t globally. This, for instance, enables us to obtain the

density part of the generic solvability independent of the winding number constraint (see [3, p. 79]). Although for the asymptotic behavior as $x \to +\infty$ we still need a different operator $Id - C_x^{\#}$, this operator is connected with $Id - C_x$ simply by a right invertible multiplier (see (7.5)). Also, for Schwartz class scattering data, the single equation (1.2) is adequate for deriving the arbitrary polynomial decay for the corresponding potentials at $x = -\infty$.

(2) The underlying spaces are less restrictive. In general the condition for the regularization (1.2) is weaker than requiring the density of the rational functions. For instance, the Hölder spaces are not separable (see [5], [7, p. 60]), but the regularization is still valid (see [10]). We remark that, for unbounded contours, the Hölder continuity of f at ∞ is required in the sense that $z^{-1}f(z^{-1})$ is Hölder continuous at 0. Note that the transform $f \mapsto z^{-1}f(z^{-1})$ commutes with the Cauchy integral operators.

(3) The operator of the algebraic system used in [3] does not have property 4 above.

It is a classical result [8] that the matrix functions with positive definite real part on a line have only zero partial indices. This result implies that the inverse scattering problem for the $n \times n$ AKNS system with $\operatorname{Re} J = \operatorname{Re} q = 0$ is always solvable [4]. We show in this paper (Theorem 9.2) that the matrix functions with positive definite real parts on the real axis and the Schwarz reflection invariant elsewhere only have zero partial indices. This applies to the systems with $J(\bar{z})^* + J(z) = 0$ and $q(x, \bar{z})^* + q(x, z) = 0$. The contours involved could be very complicated. For example, when $J(z) = z^n J$, $J + J^* = 0$ [9], the contours consist of finitely many straight lines intersecting at the origin; for the generalized sine-Gordon equation (we replace the spectral parameter z by iz in [1]), the contour consists of the real axis and the unit circle.

In §2, the required decomposing algebras $\mathbf{H}^k(\Sigma^{\pm})$ are constructed based on Sobolev spaces. We point out here that the Cauchy integral operators are under no circumstances defined componentwise. In §3, the Riemann–Hilbert problem is introduced in a narrow sense merely to meet the requirements for the study of the inverse scattering problem. However, the related singular integral operators constructed there play central roles in the study of the Riemann–Hilbert problem in the general sense as well. Section 4 gives the regularization. Several immediate results are given with short proofs. Some of these results can be generalized by means of the techniques introduced in §9.

In §5, the generalized triangular factorization is introduced in order to accommodate the systems in which $J(z)$ may have some equal diagonal entries. In §6, the inverse scattering problem is studied. We introduce the factorized scattering data which differ from the transformed scattering data in [3]. Also the augmented contours are introduced to convert the discrete scattering data into the "continuous." The resulting Riemann–Hilbert problem is completely equivalent to the original one. In §7, by means of the generalized triangular factorization, we study the Riemann–Hilbert problem for the systems in which $J(z)$ may have some equal diagonal entries. This may also help to obtain the conditions for the decay of the potentials at the nonoblique directions for the generalized wave equation and the generalized sine-Gordon equation [1]. In §8, we construct $q(x, z)$ from the fundamental solutions of the Riemann–Hilbert problem of $e^{x \operatorname{ad} J(z)}v$ for arbitrary rational diagonal matrix functions $J(z)$ when $e^{x \operatorname{ad} J(z)}v$ is in the Sobolev spaces. We see that the z dependence of q is far from arbitrary. For the class of q obtained this way, the direct problem can be worked out by virtue of the method introduced in [3].

In §9, the general theory of the Riemann–Hilbert problem is studied. Certain classical results are extended to the case of self-intersecting contours with a pair of decomposing algebras. A partial index argument is used in the proof of Theorem 9.2 to attain generality. This is useful for the problem in which we only wish to construct solutions for certain nonlinear PDEs for which the "scattering data" and the contours need not be characterized as in this paper. Certainly, the partial index argument may be avoided if we only consider the Riemann–Hilbert factorization for scattering data, since then we have the Fredholm alternative.

2. H spaces and the Cauchy operators. To avoid a flood of parentheses we follow the rule that when operator actions and matrix multiplications are mixed in a row, we do them in order from right to left. For example,

$$AbcBd \stackrel{\text{def}}{=} A(b(c(Bd))). \tag{2.1}$$

We denote by $\mathbf{M}_n$ the complex $n \times n$ matrix algebra with the inner product

$$(a, b) \stackrel{\text{def}}{=} \operatorname{tr} b^* a. \tag{2.2}$$

The corresponding norm is denoted by $|\cdot|$. An element $h \in \mathbf{M}_n$ may also denote the functions constantly equal to h on the spaces understood from the context. In the sequel, we will simply call an $\mathbf{M}_n$-valued function a matrix function.

Let $\Sigma \subset \mathbb{C}$ be a finite union of simple smooth curves that can be either closed on the $\mathbb{C}$ plane or extended to be closed on the Riemann sphere. The set of the intersections of these contours is denoted by S and assumed to be finite. Clearly Σ is smooth except at the points of S. It is readily verified that Σ admits an orientation in the sense that Σ is the positively oriented boundary for an open set Ω^+ and the negatively oriented boundary for $\Omega^- \stackrel{\text{def}}{=} \mathbb{C} \setminus (\Sigma \cup \Omega^+)$.

Clearly Ω^+ and Ω^- can only have finitely many components and Σ has two possible orientations. Any other "orientations" fail to make the Cauchy operators defined below complementary projections. The oriented Σ is still denoted by Σ while $-\Sigma$ is used to denote Σ with the opposite orientation. The symbol Σ^+ (Σ^-) is also used to denote Σ when it is viewed as the boundary of Ω^+ (Ω^-). The $\mathbf{L}^p$ norm of a function $f : \Sigma \to \mathbf{M}_n$ is defined as

$$\|f\|_p = \left(\int_\Sigma |f|^p \right)^{1/p} |dz|. \tag{2.3}$$

We will simply write $\mathbf{L}^p(\Sigma)$ for $\mathbf{L}^p(\Sigma, \mathbf{M}_n)$.

The Cauchy integral operators C_+ and C_- on $\mathbf{L}_2(\Sigma)$ are defined as

$$C_\pm f(z'') = \lim_{z' \to z''} \frac{1}{2\pi i} \int_\Sigma \frac{f(z)dz}{z - z'} \tag{2.4}$$

where the nontangential limit $z' \to z''$ is taken from Ω^+, Ω^-, respectively. We make the assumption on Σ that $C_\pm$ are bounded from $\mathbf{L}^2$ to $\mathbf{L}^2$. The recent development in the $\mathbf{L}^p$ theory of singular integral operators on curves typically allows the curves to have corners with positive angles. Since Σ is a positively oriented contour for Ω^+ and a negatively oriented contour for Ω^-, $\pm C_\pm$ can be shown to be complementary projections. A function in $\ker C_+$ has an analytic extension to Ω^- while a function in $\ker C_-$ has an analytic extension to Ω^+.

Let Γ be a piecewise smooth simple curve. We denote by $\mathbf{H}^k(\Gamma)$ all the matrix functions f such that $f^{(j)} \in \mathbf{L}^2(\Gamma)$ for all $j = 0, \cdots, k$ in the distribution sense. For the details of constructing such functions, see [3, pp. 70–71].

We make a further assumption that $\partial\Omega_\nu$ for each component Ω_ν of $\mathbb{C}\setminus\Sigma$ does not have self-intersections. $\mathbf{H}^k(\Sigma\setminus S)$ for $k>0$ is the Hilbert space of all the matrix functions f on Σ such that $f\restriction\Sigma_\nu\in\mathbf{H}^k(\Sigma_\nu)$ for every component Σ_ν of $\Sigma\setminus S$. The norm of this Hilbert space $\|f\|_{2,k}\stackrel{\text{def}}{=}(\sum_{j=0}^k\|f^{(j)}\|_2^2)^{1/2}$ will be called the $\mathbf{H}^k$ norm. We point out that for $f\in\mathbf{H}^k(\Sigma\setminus S)$, $f^{(j)}$ for $j=0,\cdots,k-1$ are continuous on Σ_ν with limits on $\partial\Sigma_\nu$ after a correction of the function on a set of measure zero. Since the Cauchy integral operators are not bounded on $\mathbf{H}^k(\Sigma\setminus S)$, we need the subspaces $\mathbf{H}^k(\Sigma^+)$ and $\mathbf{H}^k(\Sigma^-)$ for $k>0$ of $\mathbf{H}^k(\Sigma\setminus S)$. $\mathbf{H}^k(\Sigma^+)$ contains all f on Σ such that $f\restriction\partial\Omega_\nu\in\mathbf{H}^k(\partial\Omega_\nu)$ for every component Ω_ν of Ω^+ and $\mathbf{H}^k(\Sigma^-)$ such that $f\restriction\partial\Omega_\nu\in\mathbf{H}^k(\partial\Omega_\nu)$ for every component Ω_ν of Ω^-. $\mathbf{H}^k(\Sigma)\stackrel{\text{def}}{=}\mathbf{H}^k(\Sigma^+)\cap\mathbf{H}^k(\Sigma^-)$.

The $\mathbf{H}$ spaces defined above are Hilbert spaces with continuous pointwise multiplication. This follows from the fact that the norm $\|f\|_{2,k}$ dominates the norm $\sum_{j=0}^{k-1}\|f^{(j)}\|_\infty$. These spaces are Banach algebras if their norms are replaced by certain equivalent ones. For convenience, sometimes we denote $\Sigma\setminus S$, Σ, Σ^+, or Σ^- by Σ^*. For consistency, sometimes we also denote $\mathbf{L}^2(\Sigma)$ by $\mathbf{H}^0(\Sigma^*)$, but it is not an algebra under pointwise multiplication.

The rational functions are dense in $\mathbf{H}^k(\partial\Omega_\nu)$ for every component Ω_ν of $\mathbb{C}\setminus\Sigma$. In [3] this is proven for sectors. The proof can easily be generalized. We denote by R^+ (R^-) the set of all the matrix functions with their restrictions on $\partial\Omega_\nu$ rational (after a possible modification at the self-intersections) for every component Ω_ν of Ω^+ (Ω^-). Then $R=R^+\cap R^-$ is the set of all the rational matrix functions. We also denote by R_+^+ (R_-^-) the set of all the functions in R^+ (R^-) extended to be componentwise holomorphic on Ω^+ (Ω^-). $R_\pm=R\cap R_\pm^\pm$. Clearly, under this assumption, $R^+\cap\mathbf{L}^2$ ($R^-\cap\mathbf{L}^2$) is dense in $\mathbf{H}^k(\Sigma^+)$ ($\mathbf{H}^k(\Sigma^-)$). If u is a function on $\mathbb{C}\setminus\Sigma$, by u_+ and u_- we denote the limits, if they exist, of $u(z)$ as z approaches $\Sigma\setminus S$ from Ω^+ and Ω^- respectively.

PROPOSITION 2.1. *C_+, C_- are bounded from $\mathbf{H}^k(\Sigma^-)$, $\mathbf{H}^k(\Sigma^+)$ to $\mathbf{H}^k(\Sigma)$, respectively. And C_+, C_- are bounded from $\mathbf{H}^k(\Sigma^+)$, $\mathbf{H}^k(\Sigma^-)$ to $\mathbf{H}^k(\Sigma^+)$, $\mathbf{H}^k(\Sigma^-)$, respectively.*

Proof. For instance, for C_+, if $r\in R^-\cap\mathbf{L}^2$, it is easily checked that C_+r is rational and therefore belongs to $\mathbf{H}^k(\Sigma)$. The boundedness of C_+ is deduced from the integration by parts on Σ^-. Since $R^-\cap\mathbf{L}^2$ is dense in $\mathbf{H}^k(\Sigma^-)$, we conclude that C_+ is bounded from $\mathbf{H}^k(\Sigma^-)$ to $\mathbf{H}^k(\Sigma)$. We have a parallel argument for C_-. The remaining part of the proposition follows from that $C_+-C_-=Id$. The proof can also be made independent of the rational approximation to meet the need of the decomposing algebras in which the rational functions are not necessarily dense. □

A function $f\in\mathbf{H}^k(\Sigma)$ can be approximated in $\mathbf{H}^k$ norm by either functions $r^+\in R^+\cap\mathbf{L}^2$ or functions $r^-\in R^-\cap\mathbf{L}^2$. Since

$$f-(C_+r^- - C_-r^+)=C_+(f-r^-)-C_-(f-r^+), \tag{2.5}$$

f can be in fact simply approximated by the rational functions $C_+r^- - C_-r^+$. Therefore $R\cap\mathbf{L}^2$ is dense in $\mathbf{H}^k(\Sigma)$. Furthermore $R_+^+\cap\mathbf{L}^2$ is dense in $\ker C_-$ and $R_-^-\cap\mathbf{L}^2$ dense in $\ker C_+$.

PROPOSITION 2.2. *In the $\mathbf{L}^2$ space, if $f\in\ker C_+$ and g is componentwise holomorphic on Ω^- and extends continuously to the boundary of each component of Ω^-, then $fg\in\ker C_+$.*

We have a parallel result regarding $\ker C_-$.

Proof. Since g is bounded, C_+fg can be approximated in $\mathbf{L}^2$ norm by C_+r_fg where r_f is an $\mathbf{L}^2$ rational approximation of f with no poles in Ω^-. We may conclude that $C_+fg = 0$ from $C_+r_fg = 0$. □

For the uniformity of the demonstration, we assume that Σ is unbounded. We embed $\mathbf{H}^k(\Sigma^*)$ for $k \geq 0$ into a larger Hilbert space $\mathbf{H}_I^k(\Sigma^*)$ consisting of matrix functions f on Σ with the limit $f(\infty)$ at ∞ such that

$$f - f(\infty) \in \mathbf{H}^k(\Sigma^*).$$

The norm for $\mathbf{H}_I^k(\Sigma^*)$ is defined as the square root of

$$|f(\infty)|^2 + \|f - f(\infty)\|_k^2.$$

This norm is again denoted by $\|\cdot\|_k$. The norm of a bounded operator $A : \mathbf{H}_I^k(\Sigma^*) \to \mathbf{H}_I^k(\Sigma^*)$ is denoted by $\|A\|_k$.

Clearly $\mathbf{H}_I^k(\Sigma^*)$ is isomorphic to the Hilbert space direct sum of $\mathbf{H}^k(\Sigma^*)$ and $\mathbf{M}_n$. For $k > 0$, $\mathbf{H}_I^k(\Sigma^*)$ is an inverse closed Banach algebra with identity element I (see §9).

3. Riemann–Hilbert problem and the related integral operators.

DEFINITION 3.1. A matrix function v on Σ is said to be some data on Σ of smoothness degree $k \geq 1$ if v admits a factorization $v = b^{--1}b^+$ for nonsingular $b^\pm$ such that $b^\pm - I \in \mathbf{H}^k(\Sigma^\pm)$.

Let $w^+ = b^+ - I$, $w^- = I - b^-$. We call $w = (w^+, w^-)$ a factorized data of v. The set of all the data defined above is denoted by D_k and all the corresponding factorized data by FD_k.

Clearly a data v may correspond to more than one factorized data. In the following the Riemann–Hilbert problem is described in a narrow sense in terms of the factorized data. For the Riemann–Hilbert problem in a more general sense see §9.

DEFINITION 3.2. Let $w \in FD_k$, a vector $m \in \mathbf{H}_I^j(\Sigma)$ for some $j = 0, \cdots, k$ is said to be a solution of the Riemann–Hilbert problem of data w if

$$mb^\pm - m(\infty) \in \operatorname{ran} C_\pm \tag{3.1}$$

Clearly $m_\pm \overset{\text{def}}{=} mb^\pm \in \mathbf{H}_I^j(\Sigma^\pm)$. We denote by $\mathbf{m}$ the componentwise holomorphic extension of $m_\pm$. The function m is called a vanishing solution if $m(\infty) = 0$, and a fundamental solution if $m(\infty) = I$ and if $\det \mathbf{m}$ vanishes nowhere. We also call $m_\pm$ or $\mathbf{m}$ the solution of the Riemann–Hilbert problem of v or of w.

PROPOSITION 3.1. *The Riemann–Hilbert problem of v has a fundamental solution only if*

$$\frac{1}{2\pi}\int_\Sigma d \arg \det v = 0, \tag{3.2}$$

and conversely, if $m \in \mathbf{H}_I^k(\Sigma)$ with $k > 0, m(\infty) = I$ is a solution of the Riemann–Hilbert problem of v and (3.2) is fulfilled, then m is a fundamental solution. Furthermore, if $\det v = 1$, then $\det \mathbf{m} = 1$.

Proof. Suppose that v admits a fundamental solution $m_\pm$, then $\det m_\pm$ is a fundamental solution of the scalar Riemann–Hilbert problem of $\det v$. Therefore (3.2) is fulfilled.

For the converse part, let $\mathbf{m}$ be a solution that meets the hypotheses and (3.2) holds. It follows that $\det \mathbf{m}$ is a solution for the scalar Riemann–Hilbert problem of $\det v$ with $\det \mathbf{m}(\infty) = 1$. Condition (3.2) implies that the scalar Riemann–Hilbert

problem of $\det v$ admits a fundamental solution $\mathbf{m}_1$; then clearly $(\det \mathbf{m})\mathbf{m}_1^{-1}$ is an entire function with limit 1 at ∞ and therefore equals 1. The last part of the proposition also easily follows from the fact that $\mathbf{m}_1 = 1$ in this special case. □

PROPOSITION 3.2. *If the Riemann–Hilbert problem of $w \in FD_k$ admits a fundamental solution $m \in \mathbf{H}_I^j(\Sigma)$ for some $j > 0$, then it is unique in $\mathbf{L}_I^2(\Sigma)$.*

Proof. Let $n \in \mathbf{L}_I^2(\Sigma)$ also be a fundamental solution. Applying Proposition 2.2 we have

$$nm^{-1} - I = nb^{\pm}(mb^{\pm})^{-1} - I \in \ker C_+ \cap \ker C_- = 0. \qquad \square$$

The Riemann–Hilbert problem of $w \in FD_k$ is related to an operator C_w bounded from $\mathbf{H}_I^j(\Sigma)$ to $\mathbf{H}_I^j(\Sigma)$ for every $j = 0, \cdots, k$ defined as

$$(C_w\phi) = C_+\phi w^- + C_-\phi w^+.$$

PROPOSITION 3.3. *The vector $m \in \mathbf{L}_I^2(\Sigma)$ is a solution of the Riemann–Hilbert problem of w with $m(\infty) = h$ if and only if m is a solution of the system*

$$m = h + C_w m^{\times}. \tag{3.4}$$

Proof. The above system can be written as

$$m = h + C_+ m w^- + C_- m w^+. \tag{3.5}$$

Suppose that m is a solution of this system. From 3.5 clearly $m(\infty) = h$. Thus we have

$$\begin{aligned} C_+ m(w^+ + w^-) &= C_+ m w^- + C_- m w^+ - C_- m w^+ + C_+ m w^+ \\ &= m - h + m w^+ = m b^+ - m(\infty), \end{aligned} \tag{3.6}$$

and similarly

$$C_- m(w^+ + w^-) = m b^- - m(\infty). \tag{3.7}$$

Therefore 3.1 is satisfied. Conversely, suppose that m is a solution of the Riemann–Hilbert problem of w with $m(\infty) = h$, then

$$\begin{aligned} C_+ m w^- + C_- m w^+ &= C_+(m - m(\infty)) \\ &\quad - C_+(m b^- - m(\infty)) - C_-(m - m(\infty)) \\ &\quad + C_-(m b^+ - m(\infty)) = m - h. \end{aligned}$$

Therefore (3.5) is satisfied. □

The following proposition gives the equivalence relation between $Id - C_w$ and $Id - C_{w'}$ when w and w' belong to the same v.

PROPOSITION 3.4. *Suppose that w and w' are the factorized data of the same data v. Let $u = b'^+ b^{+-1} (= b'^- b^{--1} \in \mathbf{H}_I^k(\Sigma))$ and define an invertible operator U on $\mathbf{H}_I^k(\Sigma)$ as*

$$U\phi = \phi u.$$

Then we have the relation $(Id - C_w)U = Id - C_{w'}$.

Proof.

$$\begin{aligned} (Id - C_w)U\phi &= \phi u - C_+\phi u(I - b^-) + C_-\phi u(I - b^+) \\ &= \phi u - C_+\phi(u - b'^-) - C_-\phi(b'^+ - u) \\ &= \phi - C_+\phi w'^- - C_-\phi w'^+ = (Id - C_{w'})\phi. \end{aligned} \qquad \square$$

4. Regularization. For $w \in FD_k$, we define the associated factorized data $\tilde{w}$ of w as $\tilde{w}^{\pm} = \pm b^{\pm -1} \mp I$ where $b^{\pm}$ are determined by w as in Definition 3.1. Clearly $\tilde{w}$ also belongs to FD_k. Since $\tilde{w}$ is determined by w, we define operator T_w as

$$T_w\phi = C_+(C_-\phi(w^+ + w^-))\tilde{w}^- + C_-(C_+\phi(w^+ + w^-))\tilde{w}^+. \tag{4.1}$$

Clearly T_w is bounded from $\mathbf{H}_I^i(\Sigma)$ to $\mathbf{H}_I^i(\Sigma)$ for any $i = 0, \cdots, k$. It can be shown by the rational approximations of $\tilde{w}^+$ and $\tilde{w}^-$ that T_w is compact.

Remark. In general, the density of rational functions ensure the compactness of T_w. Nonetheless, as mentioned in the introduction, T_w can be compact on certain spaces without the density of the rational functions.

PROPOSITION 4.1. *$Id - C_w$ is Fredholm and*

$$Id + T_w = (Id - C_{\tilde{w}})(Id - C_w). \tag{4.2}$$

Proof. Using the fact that $\pm C_{\pm}$ are complementary projections gives

$$\begin{aligned} C_{\tilde{w}}C_w\phi &= C_+(C_w\phi)\tilde{w}^- + C_-(C_w\phi)\tilde{w}^+ \\ &= C_+(C_+\phi w^- + C_-\phi w^+)\tilde{w}^- + C_-(C_+\phi w^- + C_-\phi w^+)\tilde{w}^+ \\ &= C_+(C_-\phi(w^+ + w^-))\tilde{w}^- + C_-(C_+\phi(w^+ + w^-))\tilde{w}^+ \\ &\quad + C_+(\phi w^-)\tilde{w}^- - C_-(\phi w^+)\tilde{w}^+ \\ &= T_w\phi + C_+\phi(w^- + \tilde{w}^-) + C_-\phi(w^+ + \tilde{w}^+) \\ &= (T_w + C_{\tilde{w}} + C_w)\phi. \end{aligned}$$

Since C_w and $C_{\tilde{w}}$ are in the symmetric positions, $Id - C_{\tilde{w}}$ is also a right regulator for $Id - C_w$. Therefore the operator $Id - C_w$ is Fredholm (see [10, p. 33]). □

In the rest of this section, we assume that $w^{\pm}$ are nilpotent merely to meet the requirements for the study of the inverse scattering problem. In §9, we will see that this assumption can be replaced by (3.2).

PROPOSITION 4.2 (Fredholm alternative). *The operator $Id - C_w$ has zero Fredholm index and is invertible whenever $Id + T_w$ is.*

Proof. Since $w^{\pm}$ are nilpotent, it follows from the fact that $\det(I \pm \varsigma w^{\pm}) = 1$ that ςw is in FD_k for any complex number ς. This continuously (analytically) deforms C_w to zero. Therefore the Fredholm index of $Id - C_w$ is zero. Now if $Id + T_w$ is invertible, then $Id - C_w$ is injective and therefore invertible. □

PROPOSITION 4.3 (Analytic Fredholm alternative). *If $w = w_\varsigma$ depends on a parameter ς analytically, then either $(Id - C_{w_\varsigma})^{-1}$ is meromorphic in ς or $Id - C_{w_\varsigma}$ is invertible for no ς.*

Proof. The corresponding $\tilde{w}_\varsigma$ is clearly also analytic in ς. Now suppose that $Id - C_{w_\varsigma}$ is invertible for $\varsigma = \varsigma_0$. Then there is a compact operator B such that $Id - C_{\tilde{w}_\varsigma} + B$ is injective for $\varsigma = \varsigma_0$ (see [10, pp. 34–38]). Therefore

$$(Id - C_{\tilde{w}_\varsigma} + B)(Id - C_{w_\varsigma}) = Id + T_{w_\varsigma} + B(Id - C_{w_\varsigma})$$

is injective for $\varsigma = \varsigma_0$ and so its inverse is meromorphic in ς. It follows from Proposition 4.2 that

$$(Id - C_{w_\varsigma})^{-1} = (Id + T_{w_\varsigma} + B(Id - C_{w_\varsigma}))^{-1}(Id - C_{\tilde{w}_\varsigma} + B)$$

whenever the right-hand side exists. Therefore $(Id - C_{w_\varsigma})^{-1}$ is meromorphic in ς.

Combining Proposition 4.2 with Proposition 3.2 gives Proposition 4.4. □

PROPOSITION 4.4. *If the Riemann–Hilbert problem of $w \in FD_k$ admits a fundamental solution $m \in \mathbf{H}_I^j(\Sigma)$ for some $j > 0$, then $Id - C_w$ is invertible on $\mathbf{H}_I^j(\Sigma)$ for all $j = 0, \cdots, k$.*

PROPOSITION 4.5. *Suppose that $w \in FD_k$. If $Id - C_w$ on space $\mathbf{H}_I^j(\Sigma)$ is invertible for any $j \leq k$, then it is invertible for all $j \leq k$.*

Proof. Suppose that $Id - C_w$ is invertible on $\mathbf{L}_I^2(\Sigma)$, then $\ker(Id - C_w) = 0$ in $\mathbf{L}_I^2(\Sigma)$ and so in $\mathbf{H}_I^j(\Sigma)$ for any $j = 0, \cdots, k$. Therefore by the Fredholm alternative, $Id - C_w$ is invertible on all these spaces.

Now suppose that $Id - C_w$ is invertible on $\mathbf{H}_I^j(\Sigma)$ for some j with $1 \leq j \leq k$; then $m = (Id - C_w)^{-1} I$ is a fundamental solution for the Riemann–Hilbert problem in $\mathbf{H}_I^j(\Sigma)$. Therefore $Id - C_w$ is invertible on $\mathbf{L}_I^2(\Sigma)$. □

5. Generalized triangular factorization. The study of the inverse scattering problem for system (1.1) (in which J may have some equal diagonal entries) requires the following notation.

For $A \in \mathbf{M}_n$, the operator $\operatorname{ad} A$ on $\mathbf{M}_n$ is defined as $\operatorname{ad} A(B) = [A, B] \stackrel{\text{def}}{=} AB - BA$. Let $A \in \mathbf{M}_n$ be real and diagonal, then $\operatorname{ad} A$ is a self-adjoint operator on $\mathbf{M}_n$. For $B \in \mathbf{M}_n$, we call $\chi_{(-\infty,0)}(\operatorname{ad} A)(B)$, $\chi_{(0,+\infty)}(\operatorname{ad} A)(B)$, and $\chi_{\{0\}}(\operatorname{ad} A)(B)$ the A lower triangular, A upper triangular, and A diagonal part of B, respectively. The phrases such as "A (off) diagonal," "A (strictly) upper triangular," and "A (strictly) lower triangular" are understood.

To visualize the above terminologies, we permute the basis of $\mathbf{M}_n$, such that the diagonal entries of A are reordered nonincreasingly. Then A has the form

$$A = \begin{pmatrix} a_1 I_1 & 0 & \dots & 0 \\ 0 & a_2 I_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_k I_k \end{pmatrix}, \tag{5.1}$$

where $a_1, \cdots, a_k$ are distinct and $I_1, \cdots, I_k$ are some identity blocks. Accordingly, any matrix $B \in \mathbf{M}_n$ can be written as

$$B = \begin{pmatrix} B_{11} & B_{12} & \dots & B_{1n} \\ B_{21} & B_{22} & \dots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \dots & B_{nn} \end{pmatrix}, \tag{5.2}$$

where B_{ij}, $1 \leq i, j \leq n$, are blocks such that the orders of $B_{11}, B_{22}, \cdots, B_{nn}$ are equal to those of $I_1, I_2, \cdots, I_n$, respectively. Under this expression,

$$\chi_{(-\infty,0)}(\operatorname{ad} A)(B) = \begin{pmatrix} 0 & 0 & \dots & 0 \\ B_{21} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \dots & 0 \end{pmatrix}, \tag{5.3}$$

$$\chi_{(0,+\infty)}(\operatorname{ad} A)(B) = \begin{pmatrix} 0 & B_{12} & \dots & B_{1n} \\ 0 & 0 & \dots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \tag{5.4}$$

and

$$\chi_{\{0\}}(\operatorname{ad} A)(B) = \begin{pmatrix} B_{11} & 0 & \dots & 0 \\ 0 & B_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B_{nn} \end{pmatrix}. \tag{5.5}$$

We also call

$$d^k(B) \stackrel{\text{def}}{=} \det \begin{pmatrix} B_{11} & B_{12} & \dots & B_{1k} \\ B_{21} & B_{22} & \dots & B_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \dots & B_{kk} \end{pmatrix}, \qquad 1 \le k \le n \tag{5.6}$$

the A upper principal minors of B, and respectively

$$d_k(B) \stackrel{\text{def}}{=} \det \begin{pmatrix} B_{kk} & B_{k,k+1} & \dots & B_{kn} \\ B_{k+1,k} & B_{k+1,k+1} & \dots & B_{k+1,n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{nk} & B_{n,k+1} & \dots & B_{nn} \end{pmatrix}, \qquad 1 \le k \le n \tag{5.7}$$

the A lower principal minors of B. Note that these A principal minors are just a part of the lower and upper principal minors of B in the usual sense. It is easy to check that we have the following proposition for the generalized triangular factorization.

PROPOSITION 5.1. *If no A upper principal minors of $v \in \mathbf{M}_n$ vanish, then v admits a unique triangular factorization*

$$v = b^{-1}\Delta a \tag{5.8}$$

such that $b - I$ is strictly A lower triangular, $a - I$ strictly A upper triangular, and Δ A diagonal.

6. The inverse scattering problem. Let $J = \operatorname{diag}(\lambda_1(z), \cdots, \lambda_n(z))$ be a rational matrix function with its constant term being zero. We assume that $J(z)$ and $\operatorname{ad} J(z)$ have exactly the same poles on the Riemann sphere and denote the set of these poles by P_J.

We define $\Sigma_0 \subset \mathbb{C}$ as the closure of

$$\{z \in \mathbb{C} | \operatorname{Re}(\lambda_j(z) - \lambda_k(z)) = 0, \quad \text{some } \lambda_j \neq \lambda_k\}. \tag{6.1}$$

and denote by S the set of all the self-intersections of Σ_0. Let $\Omega_0^{\pm}$ be the regions corresponding to Σ_0. Clearly, all the poles of $J(z)$ lie on $\Sigma \cup \{\infty\}$. Since $J(z)$ is fixed, the phrases such as "$\operatorname{ad}(\operatorname{Re} J(z))$ diagonal," and so forth, will be simply replaced by "z diagonal," etc. The spectrum projection operators $\chi_{(-\infty,0)}(\operatorname{ad}(\operatorname{Re} J(z)))$, $\chi_{(0,+\infty)}(\operatorname{ad}(\operatorname{Re} J(z)))$, and $\chi_{\{0\}}(\operatorname{ad}(\operatorname{Re} J(z)))$ are constant in each component of $\mathbb{C} \setminus \Sigma_0$. For $z \in \Sigma_0 \setminus S$ by z^+, we mean a point in Ω_0^+ near z and by z^- a point in Ω_0^- near z. Thus we can use phrases such as "z^+ upper triangular" and "z^- upper triangular," etc. A simple fact we will use is that, if for $z \in \Sigma_0 \setminus S$, a is z^- lower (upper) triangular and z diagonal, then it is z^+ upper (lower) triangular. This follows simply from the fact that $\operatorname{Re}(\lambda_i(z) - \lambda_j(z))$ is harmonic in $(\operatorname{Re} z, \operatorname{Im} z)$, and so it changes sign across the curve $\{\operatorname{Re}(\lambda_i(z) - \lambda_j(z)) = 0\}$.

At this stage, we make the assumption that all the diagonal entries of $J(z)$ are distinct. In the next section, we will consider the possibility of removing this restriction. The data for the Riemann–Hilbert problem for inverse scattering depend on a parameter x as $v_x \stackrel{\text{def}}{=} e^{x \operatorname{ad} J(z)} v$.

Let D be a finite subset of $\mathbb{C}$ disjoint from Σ_0 (see the remarks following Definition 6.3). The continuous parts of the scattering data, denoted by v_c, are supported on Σ_0 and the discrete parts of the scattering data, denoted by v_d, are supported on D. In the following the scattering data are characterized by five conditions. The first four are essentially based on the characterizations for the scattering data given in [3]. The last one is designed to ensure that $e^{x \operatorname{ad} J(z)} v$ is in the right Sobolev spaces.

DEFINITION 6.1. The discrete part and the continuous part of the scattering data of smoothness degree $k \geq 1$ are defined as follows:

(1) $v_c - I \in \mathbf{H}^k(\Sigma \setminus S)$, and $v_d(z_i)$ is strictly z_i upper triangular with a single nonzero entry for each $z_i \in D$;

(2) For each $z' \in S \setminus P_J$, there is a componentwise polynomial matrix function $a_{k,z'}(z - z')$ on $\mathbb{C} \setminus \Sigma$ such that $a_{k,z'}(z - z') - I$ is z strictly upper triangular and such that near $z = z'$

$$v_c(z) = (a^-_{k,z'}(z - z'))^{-1} a^+_{k,z'}(z - z') + o(|z - z'|^{k-1}) \tag{6.2}$$

(3) $v_c(z)$ is z diagonal and all the z^- lower principal minors are 1;

(4) A winding number constraint about the z^- upper principal minors of $v_c(z)$ and v_d. Since the winding number constraint will not be explicitly used in this paper we do not describe it here. For details see [3].

(5) If $z = \infty$ is a pole of $J(z)$ with multiplicity n, then

$$z^{j(n-1)}(v_c - I)^{(k-j)} \in \mathbf{L}^2 \quad \text{for } j = 1, \cdots, k; \tag{6.3}$$

if $z = z_0$ is a finite pole of $J(z)$ with multiplicity n, then

$$(z - z_0)^{-j(n+1)}(v_c - I)^{(k-j)} \in \mathbf{L}^2 \quad \text{for } j = 1, \cdots, k. \tag{6.4}$$

DEFINITION 6.2. Let D be covered by $|\Lambda|$ disjoint closed disks that do not intersect with Σ_0 such that each $z_i \in D$ is the center of exactly one of these disks. We add the boundary circles of these disks to form an augmented contour Σ. Since these added circles do not intersect with Σ_0, we may keep the original orientation on the Σ_0 part of Σ and clearly this determines the orientation uniquely for Σ.

DEFINITION 6.3. Now we define the scattering data v supported on Σ as follows:

(1) $v = v_c$ on Σ_0;

(2) On the added circle surrounding $z_i \in D \cap \Omega_0^{\pm}$, $v = I \pm v_d(z_i)/(z - z_i)$.

We denote by SD_k the set of all scattering data on Σ defined above.

Condition (6.2) guarantees that $v \in D_k$ (see Proposition 6.1 below). Conditions (6.3) and (6.4) guarantee $v_x \overset{\text{def}}{=} e^{x \operatorname{ad} J(z)} v \in D_k$ for every $x \in \mathbb{R}$.

Let $\mathbf{m}(x, z)$ be the fundamental solution of the Riemann–Hilbert problem of v_x. Using the fact $v_d(z_i)^2 = 0$, we see that $\mathbf{m}(x, z)$, from outside the added circle surrounding $z_i \in D$, extends meromorphically to inside the circle as

$$\mathbf{m}_i \overset{\text{def}}{=} \mathbf{m}\left(I + \frac{e^{x \operatorname{ad} J(z)} v_d(z_i)}{z - z_i}\right).$$

Using $v_d(z_i)^2 = 0$ and $(\operatorname{ad} J'(z_i) v_d(z_i)) v_d(z_i) = 0$ we obtain

$$\mathbf{m}_{i,-1} = \mathbf{m}_{i,0} e^{x \operatorname{ad} J(z_i)} v_d(z_i)$$

where $\mathbf{m}_{i,-1}$ ($\mathbf{m}_{i,0}$) is the -1st (0th) Laurent coefficient of $\mathbf{m}_i$ at z_i. This relation is used in [3] to describe the Riemann–Hilbert problem with discrete spectrum.

For the inverse scattering problem, it often happens that, on a portion of Σ_0, only certain columns of $\mathbf{m}$ commit the jump. Therefore the remaining columns may be permitted to have poles. This gives rise to the discrete scattering data on the contour Σ_0. In this case, the added circles may intersect with Σ_0. But if $v_d(z_i)(v(z) - I) = (v(z) - I) v_d(z_i) = 0$ for z in a neighborhood of z_i, we may still define the scattering data on the augmented contour in a similar fashion as in Definition 6.3.

DEFINITION 6.4. A factorized data $w \in FD_k$ of $v \in SD_k$ is said to be a factorized scattering data if

(1) $w^-(z)$ is z^- strictly upper triangular and $w^+(z)$ z^+ strictly upper triangular;

(2) $w^\pm(z)$ are z diagonal in an open set containing P_J.

The set of all the factorized scattering data obtained above is denoted by FSD_k.

PROPOSITION 6.1. *Every scattering data $v \in FD_k$ admits a factorized scattering data w.*

Proof. We cover $S \setminus P_J$ by disjoint small disks in the way that each disk contains exactly one $z' \in S \setminus P_J$. Since all the z^- lower principal minors of v_c are equal to 1, v_c admits a triangular factorization $v_c = b_1^{--1} b_1^+$ such that $b_1^+(z) - I$, $(b_1^-(z) - I)$ is z^- strictly lower (upper) triangular. Since $v_c(z)$ is z diagonal, $b_1^\pm(z)$ are also z diagonal and therefore $b_1^+(z)$ is z^+ upper triangular. Clearly $b_1^\pm - I \in \mathbf{H}^k(\Sigma_0 \setminus S)$. Define $f_{z'}(z)$ on Σ_0 near each $z' \in S \setminus P_J$ such that on each curve of Σ_0 ending at z', $f_{z'}(z)$ equals the Taylor polynomial of degree $k-1$ for $a_{k,z'}^- b_1^{--1}$ at z' which also equals, by condition 2 in Definition 6.1, the Taylor polynomial of degree $k-1$ for $a_{k,z'}^+ b_1^{+-1}$ at z' from the very curve. It follows that $f_{z'}(z) - I$ is z^+ and z^- strictly upper triangular and therefore z strictly upper triangular. Clearly there exists $f \in \mathbf{H}_I^k(\Sigma_0 \setminus S)$ satisfying:

(1) $f = I$ outside these disks;

(2) $f(z) - I$ is z strictly upper triangular;

(3) For each $z' \in S \setminus P_J$, f agrees with $f_{z'}$ in a neighborhood of z'.

It is easily checked that $f b_1^\pm \in \mathbf{H}_I(\Sigma_0^\pm)$ from conditions (2) and (5) in Definition 6.1. Now we define $b^\pm = f b_1^\pm$ on Σ_0 and $b^+ = v$, $b^- = I$ on the added circles, then clearly $b^\pm$ give rise to the desired factorized scattering data. □

We write $w_x = (e^{x \operatorname{ad} J(z)} w^+, e^{x \operatorname{ad} J(z)} w^-)$.

We remark that the factorized scattering data agrees with the transformed scattering data defined in [3] in an open set containing P_J. Therefore it is z diagonal in an open set containing P_J. This is very important because then the operator C_{w_x} is well defined for all $x \in \mathbb{R}$. Also when $v - I$ has a support away from an open set containing P_J, w has such a support as well. This preserves the analyticity in x.

THEOREM 6.1. *If w is a factorized scattering data, $\|T_{w_x}\|_0 \to 0$ as $x \to -\infty$.*

Proof. First we approximate $\tilde{w}^-$ in $\mathbf{L}^\infty$ norm by a function $u \in R^- \cap \mathbf{L}^2$, where $u(z)$ is chosen to be z strictly upper triangular with only simple poles $\{\zeta_i\}$. We denote the residue of u at ζ_i by u_{ζ_i}. Then $C_+(C_-\phi e^{x \operatorname{ad} J(z)}(w^+ + w^-))e^{x \operatorname{ad} J(z)}\tilde{w}^-$ is approximated in $\mathbf{L}^2$ norm by $C_+(C_-\phi e^{x \operatorname{ad} J(z)}(w^+ + w^-))e^{x \operatorname{ad} J(z)} u$ uniformly for $\|\phi\|_0 \le 1$. The latter equals

$$\sum_{\zeta_i \in \Omega^-} (C_{\zeta_i} \phi e^{x \operatorname{ad} J(z)}(w^+ + w^-)) \frac{e^{x \operatorname{ad} J(\zeta_i)} u_{\zeta_i}}{z - \zeta_i}.$$

Since u_{ζ_i} is strictly ζ_i upper triangular, (6.5) approaches zero in $\mathbf{L}^2$ as $x \to -\infty$ uniformly for $\|\phi\|_0 \le 1$.

For $C_-(C_+\phi e^{x \operatorname{ad} J(z)}(w^+ + w^-))e^{x \operatorname{ad} J(z)}\tilde{w}^+$, a similar argument works. □

Near $x = -\infty$, $(Id - C_{w_x})^{-1}$ can be calculated through the expansion

$$(Id - C_{w_x})^{-1} = \sum_{k=0}^{\infty} (-1)^k T_{\tilde{w}_x}^k (Id - C_{\tilde{w}_x}) \tag{6.5}$$

with the estimate $\|T_{\tilde{w}_x}^k (Id - C_{\tilde{w}_x})\|_0 = O(\|T_{w_x}\|_0^k)$.

THEOREM 6.2. *If $v \in FD_k$ and $v - I$ is supported away from a neighborhood of P_J, then for some transformed scattering data w of v, $(Id - C_{w_x})^{-1}$ is meromorphic*

in $x \in \mathbb{C}$, viewed as a function from $\mathbb{C}$ to the space of all the bounded operators on $\mathbf{H}_I^j(\Sigma)$ for $j = 0, \cdots, k$.

Proof. Clearly w can be made supported away from an open set containing P_J. Therefore w_x is entire in x. It follows from Theorem 6.1 and Proposition 4.5 that $(Id - C_{w_x})^{-1}$ is meromorphic in x. □

THEOREM 6.3. (Generic solvability). *SD_k^o is an open and dense subset of SD_k, where SD_k^o is the set of all the scattering data $v \in SD_k$ such that the corresponding Riemann–Hilbert problem admits a fundamental solution for every $x \in \mathbb{R}$.*

Proof. If $v \in SD_k^o$, and w is some factorized scattering data of v, then $Id - C_{w_x}$ is invertible for all $x \in \mathbb{R}$. By Proposition 4.5, to study the invertibility of $Id - C_{w_x}$, we only need to work on $\mathbf{L}_I^2(\Sigma)$. Note that w can be chosen in a way that it varies continuously with respect to v. If K is a compact subset of $\mathbb{R}$, by the continuity of C_{w_x} in x it must be still invertible for all $x \in K$ when w varies slightly. For x near $-\infty$, $\|T_{w_x}\|_0 < 1$. This inequality holds when w varies slightly. This says that $Id - C_{w_x}$ is invertible in a neighborhood of $x = -\infty$ and an open set containing w. Condition (4) in Definition 6.1 implies that the Riemann–Hilbert problem admits a solution in $\mathbf{L}_I^2(\Sigma)$ in an open set containing $x = +\infty$ and an open set containing w, and so implies the invertibility of $Id - C_{w_x}$ in an open set containing $x = +\infty$ and an open set containing w. Therefore we may conclude that SD_k^o is open in SD_k.

For the density part of the theorem, we assume that $w \in FSD_k$ is supported away from an open set containing P_J. By Theorem 6.2 on almost all the lines parallel to real axis $Id - C_{w_x}$ is invertible. This is tantamount to saying that for almost all the purely imaginary α, $Id - C_{(w_\alpha)_x}$ is invertible for all $x \in \mathbb{R}$. Where $w_\alpha \stackrel{\text{def}}{=} (e^{\alpha \operatorname{ad} J(z)} w^+, e^{\alpha \operatorname{ad} J(z)} w^-)$ is again a factorized scattering data. This proves the density part of the theorem because α can be chosen to be arbitrarily small. □

7. $J(z)$ with some equal diagonal entries. Now we consider the inverse problem with J which may have some equal diagonal entries. Basically everything still works with the generalized triangular factorization. In this case condition (3) in Definition 6.1 is replaced by:

$v_c(z)$ is z diagonal and admits a factorization $v_c = b^{--1} b^+$ such that $b^-(z) - I$ is z^- strictly lower triangular and $b^+(z) - I$ z^- strictly upper triangular (and therefore z^+ strictly upper triangular).

Note that this change is consistent with the direct problem. Set $z_0 \notin \Sigma$. It is easily seen that all the results obtained in §6 are still valid, except condition (4) in Definition 6.1 cannot be simply described in terms of the winding numbers of certain principal minors. This is because in condition (4) a Riemann–Hilbert problem of a z_0 diagonal matrix function (it is not necessarily diagonal) is involved. Since this problem and, more generally, the problem of discrete scattering data describing all possible multiple poles, has a great deal to do with the direct problem as well, we will study it in a different paper. In the following, we only consider a special case which assumes that $\Sigma = \Sigma_0$ and that the z^- upper principal minors of $v_c(z)$ do not vanish. By Proposition 5.1, v admits a factorization $v = b^{--1} \Delta b^+$ such that $b^+(z) - I$ is z^+ strictly lower triangular and $b^-(z) - I$ strictly z^- lower triangular, and $\Delta(z)$ is z_0 diagonal for a $z_0 \notin \Sigma$. Since $v(z)$ is z diagonal, $b^\pm(z)$ must also be z diagonal by the uniqueness of the triangular factorization. It is consistent with the direct problem to assume a statement similar to condition (2) in Definition 6.1 with the modification that $a_{k,z'}(z - z')$ is z lower triangular and nonsingular at $z = z'$. We write (6.2) down

again

$$v(z) = (a^-_{k,z'}(z - z'))^{-1} a^+_{k,z'}(z - z') + o(|z - z'|^{k-1}). \tag{7.1}$$

Let $b^{\pm}_{k,z'}(z - z')$ be the z_0 diagonal part of $a^{\pm}_{k,z'}(z - z')$. Then by the uniqueness of the triangular factorization,

$$\Delta(z) = (b^-_{k,z'}(z - z'))^{-1} b^+_{k,z'}(z - z') + o(|z - z'|^{k-1}). \tag{7.2}$$

It can be shown by virtue of the proof for Proposition 6.1, $\Delta(z) \in D_k$. It is consistent with the direct problem to assume that the Riemann–Hilbert problem of Δ admits the fundamental solution $\delta_\pm$ (see the "winding number constraint" in [3]), we have

$$\Delta = \delta_-^{-1} \delta_+. \tag{7.3}$$

$\delta_\pm$ are z_0 diagonal because Δ is. Define $v^\# = \delta_- v \delta_+^{-1}$ and $b^{\#\pm} = \delta_\pm b^\pm \delta_\pm^{-1}$. Then $v^\# = b^{\#--1} b^{\#+}$. It is easily checked that $v^\#$ also satisfies condition (2) in Definition 6.1 with the modification that $a_{k,z'}(z - z') - I$ is z strictly lower triangular. $v^\#$ satisfies all the conditions in Definition 6.1 with the upper and the lower triangularities switched. Therefore near $x = +\infty$, $e^{x \operatorname{ad} J(z)} v^\#$ admits a Riemann–Hilbert factorization

$$e^{x \operatorname{ad} J(z)} v^\# = m_-^{\#-1} m_+^\#. \tag{7.4}$$

Since $\delta_\pm$ is z_0 diagonal, $m_\pm \overset{\text{def}}{=} m^\#_\pm \delta_\pm$ is the solution for $e^{x \operatorname{ad} J(z)} v$. It is easily seen from system (1.1) that the potential constructed from $v^\#$ equals that from v.

Remark. Now let w, $w^\#$, and w_0 be certain factorized data of v, $v^\#$, and Δ, respectively, then (see (9.9))

$$Id - C_w = (Id - C_{w^\#}) U^\# (Id - C_{w_0}) U^{-1}, \tag{7.5}$$

where U and $U^\#$ are the invertible multipliers for corrections of the factorizations of v and $v^\#$ respectively as in Proposition 3.4. By virtue of the proof of Proposition 4.1, $Id - C_{w_0}$ is invertible because the Riemann–Hilbert problem of Δ admits the fundamental solution $\delta_\pm$. Therefore $Id - C_w$ and $Id - C_{w^\#}$ are connected by a right invertible multiplier.

Note in the above that Δ in general is not diagonal; therefore the Riemann–Hilbert problem is not a scalar problem. In §9, we will see that for the inverse problem for certain systems the Riemann–Hilbert problem of Δ always admits a fundamental solution.

The above generalization certainly has some effects on the direct problem as well. For example, the potentials of the $n \times n$ AKNS system need to be J off diagonal and clearly this is more restrictive than being off diagonal.

8. Reconstruction of the potentials. Let $m(x, \cdot)$ be a fundamental solution of w_x. We ask what kind of potentials $q(x, z)$ can be constructed such that

$$\mathbf{m}'(x, z) - \operatorname{ad} J(z) \mathbf{m}(x, z) = q(x, z) \mathbf{m}(x, z). \tag{8.1}$$

Clearly if such a $q(x, z)$ exists, it is uniquely determined by $\mathbf{m}$ because $\mathbf{m}$ is nonsingular. Let $q(x, z)$ be an undetermined rational function in z with the same poles as $J(z)$ in the way that if $z = \infty$ is a pole of $J(z)$ of the multiplicity n, then it is a pole of $q(x, z)$ of multiplicity $n - 1$; and if $z = z_0$ is a finite pole of $J(z)$ of multiplicity n, then it is also a pole of $q(x, z)$ of multiplicity n. Now we formally define an operator L on functions of variable $(x, z) \in \mathbb{R} \times \Sigma$ as

$$L\phi(x, z) = \left(\frac{d}{dx} - \operatorname{ad} J(z) - q(x, z) \right) \phi(x, z). \tag{8.2}$$

Let m be a fundamental solution of the Riemann–Hilbert problem of w; a formal calculation gives

$$(Id - C_{w_x})Lm = (Id - C_{w_x})Lm - L(Id - C_{w_x})m - q = [L, C_{w_x}]m - q. \tag{8.3}$$

To make the above calculation rigorous, let $w \in FSD_1$, and if ∞ is a pole of $J(z)$ with multiplicity n, assume $z^{2n-1}w^{\pm} \in \mathbf{L}^2(\Sigma)$, and if z_0 is a finite pole of $J(z)$ with multiplicity n, assume $(z - z_0)^{-(2n+1)}v \in \mathbf{L}^2(\Sigma)$. Under these assumptions, $\operatorname{ad} Jw \in FSD_1$, then since $(d/dx)C_{w_x} = C_{\operatorname{ad} Jw_x}$, $(d/dx)m(x, \cdot) \in \mathbf{L}^2(\Sigma)$. Also $C_{w_x}Lm \in \mathbf{L}^2$. Therefore the above calculation can be carried out rigorously. If

$$[L, C_{w_x}]m - q = 0, \tag{8.4}$$

then $Lm = C_{w_x}Lm \in \mathbf{L}^2$. Since the Riemann–Hilbert problem admits a fundamental solution, $Id - C_{w_x}$ is injective (invertible). It follows that $Lm = 0$. It is easily checked that this extends to (8.1). Therefore (8.4) is the formula for the reconstruction of q. Expression (8.4) may be written explicitly as

$$\int_\Sigma \left(\frac{\operatorname{ad} J(\zeta) - \operatorname{ad} J(z)}{\zeta - z} m(x, \zeta) + \frac{q(x, \zeta) - q(x, z)}{\zeta - z} m(x, \zeta) \right) d\mu_x(\zeta) = q(x, z), \tag{8.5}$$

with $d\mu_x(z) = (2\pi i)^{-1} e^{x \operatorname{ad} J(z)} w(z) dz$.

Example 8.1. ($n \times n$ AKNS) $J(z) = zJ$, $q(x, z) = q(x)$.

$$q(x) = \operatorname{ad} J \int_\Sigma m(x, \zeta) d\mu_x(\zeta). \tag{8.6}$$

Therefore in the direct problem we must assume that q is J off diagonal.

Example 8.2. (For Landau–Lifshitz equation [11]). Let $J(z) = z^{-1}J$, $q(x, z) = z^{-1}q(x)$.

$$-q = \operatorname{ad} J \int_\Sigma \zeta^{-1} m(x, \zeta) d\mu_x(\zeta) + q \int_\Sigma \zeta^{-1} m(x, \zeta) d\mu_x(\zeta). \tag{8.7}$$

This can be written as

$$q(x, z) + J(z) = \left(\int_\Sigma \zeta^{-1} m(x, \zeta) d\mu_x(\zeta) + I \right) J \left(\int_\Sigma \zeta^{-1} m(x, \zeta) d\mu_x(\zeta) + I \right)^{-1} \tag{8.8}$$

Therefore in this case in the direct problem we must assume that $q + J$ is similar to J.

In (8.8), $\int_\Sigma \zeta^{-1} m(x, \zeta) d\mu_x(\zeta) + I$ is nonsingular because it is $m(x, 0)$. In general, q is always solvable from (8.4), which determines the basic algebraic structures of the potentials. We point out that this is consistent with the gauge transform in the direct problem in order to acquire the desired decay at the poles of $J(z)$ for the scattering data.

PROPOSITION 8.1. *For the $n \times n$* AKNS *system, it is given in* [3] *that near $x = -\infty$,*

$$\|C_{w_x} I\| = O(x^{-k-1}). \tag{8.9}$$

Let m be a fundamental solution of w, then

$$\|m(x, \cdot) - I\|_0 = O(x^{-k-1}) \quad \text{as } x \to -\infty. \tag{8.10}$$

Proof. Since

$$(Id - C_{w_x})^{-1} = (Id - T_{w_x})^{-1}(Id - C_{\tilde{w}_x}) \quad \text{for } x \text{ near } -\infty\,, \tag{8.11}$$

$$\|(Id - C_{w_x})^{-1}\|_0 \leq c \quad \text{for } x \text{ near } -\infty , \tag{8.12}$$

for some constant c independent of x. It follows

$$\|m(x,\cdot) - I\|_0 = \|(Id - C_{w_x})^{-1} C_{w_x} I\|_0 \leq c\|C_{w_x} I\|_0 = O(x^{-k-1}). \tag{8.13}$$

The desired decay of the potential q at $x = -\infty$ can be derived by using

$$q = \frac{\operatorname{ad} J}{2\pi i}\left(\int_{\Sigma^+} m e^{x \operatorname{ad} J(z)} w^+ + \int_{\Sigma^-} m e^{x \operatorname{ad} J(z)} w^-\right). \tag{8.14}$$

At $x = +\infty$, we consider the problem normalized at $x = +\infty$ (see [3, p. 79]). For general $J(z)$, to obtain the desired decay for q, we need to study the asymptotic behaviors of the oscillatory integral $C_{w_x} I$ and its Laurent expansion at the poles of $J(z)$. We remark here that if $v_c = 0$, then by contour integration, C_{w_x} reduces to an algebraic operator and the integrals in (8.6), (8.8), and (8.14) to summations.

We conclude this section by a proposition often seen in the inverse scattering problem.

PROPOSITION 8.2. $\operatorname{tr} q = 0$.

Proof. $Lm = 0$ gives

$$q = m' m^{-1} + (J - mJm^{-1}). \tag{8.15}$$

Therefore

$$\operatorname{tr} q = \operatorname{tr} m' m^{-1}. \tag{8.16}$$

The estimate near P_J yields $\operatorname{tr} q = 0$. □

9. Further results in the Riemann–Hilbert problem and their applications to the inverse scattering problem. In this section, we work on a pair of decomposing algebras (defined below) in place of $\mathbf{H}_J^k(\Sigma^\pm)$.

Let Σ be a contour in the Riemann sphere passing ∞ and $C(\Sigma^\pm)$ be the algebra of all scalar functions on Σ with their restrictions on the boundary of each component of $\Omega^\pm$ being continuous (after a modification at the self-intersections). Letting $B \subset C(\Sigma^\pm)$ be a Banach algebra, we denote by B^0 the maximal ideal of B consisting of all functions in B vanishing at ∞. The Banach algebra $B \subset C(\Sigma^\pm)$ is said to be

(1) Inverse closed if the fact that $\phi \in B$ is invertible in $C(\Sigma^\pm)$ implies that ϕ is invertible in B;

(2) Componentwise independent if for every vector $\phi \in B$ and every component Ω_ν of $\Omega^\pm$, the function ϕ_ν equal to ϕ on $\partial\Omega_\nu$ and zero elsewhere is in B;

(3) Decomposing if $C_\pm$ are bounded on B_0.

We write $B_n = \mathbf{M}_n \otimes B$ and denote by GB_n the general linear group of B_n and by SB_n the special linear group of B_n.

If B is a decomposing algebra, we denote by $\mathcal{B}_n$ the space of componentwise holomorphic matrix functions $\mathbf{m}$ on Ω with their boundary values $m_\pm \in m_\pm(\infty) + C_\pm B_n^0$ by $G\mathcal{B}_n$, the general linear group of $\mathcal{B}_n$. The following theorem for a pair of decomposing algebras on a self-intersecting contour Σ is a generalization of the classical Riemann–Hilbert factorization theorem in a single decomposing algebra.

THEOREM 9.1. *Let $E(\Sigma^\pm) \subset C(\Sigma^\pm)$ be a pair of inverse closed componentwise independent decomposing algebras satisfying*

(1) $R^\pm \subset E(\Sigma^\pm) \subset \mathbf{L}_J^2(\Sigma)$,

(2) $C_+E^0(\Sigma^-), C_-E^0(\Sigma^+) \subset E(\Sigma) \overset{\text{def}}{=} E(\Sigma^+) \cap E(\Sigma^-)$,

(3) *For every $a_\pm \in E(\Sigma^\pm)$, the operator $\phi \mapsto C_\pm a_\pm C_\mp \phi$ on $E^0(\Sigma^+)$ and $E^0(\Sigma^-)$ is compact,*

then every $v \in GE_n(\Sigma^-) \cdot GE_n(\Sigma^+)$ *admits a Riemann–Hilbert factorization*

$$v = m_- \theta m_+ \tag{9.1}$$

where $\mathbf{m} \in G\mathcal{E}_n(\Sigma)$ *and* θ *is as in* (9.2) *below with the integers* k_i *uniquely determined by* v *up to a permutation. These integers are called the partial indices of* v.

Furthermore, if in addition, $\det v = 1$, *then* $\mathbf{m}$ *can be chosen such that* $\det \mathbf{m} = 1$ *and therefore* $v \in SE_n(\Sigma^-) \cdot SE_n(\Sigma^+)$.

Proof. Our proof is based on a standard induction over the number of components of $\mathbb{C} \setminus \Sigma$.

We first assume that $v \in GE_n(\Sigma^-)$. Let Ω' be a component of Ω^-. By the assumption that $\partial\Omega'$ does not have self-intersections, if $\Sigma = \partial\Omega'$, then we are done according to the results in [7]. Otherwise we write $\Sigma_2 = \partial\Omega'$, $\Sigma_1 = \overline{\Sigma \setminus \Sigma_2}$ and $\Omega_i = \bar{\mathbb{C}} \setminus \Sigma_i$ for $i = 1, 2$. We have the disjoint unions

$$\Omega^- = \Omega_1^- \cup \Omega_2^-, \qquad \Omega_1^+ = \Omega_2^- \cup \Omega^+,$$

where none of Ω_1^-, Ω_2^-, or Ω^+ is empty. Fix $z_+ \in \Omega^+$, $z_- \in \Omega_1^-$ and $z'_- \in \Omega_2^-$.

We write $v_i = v \restriction \Sigma_i$ for $1 = 1, 2$. Then $v_1 \in GE_n(\Sigma_1^-)$ where $E_n(\Sigma_1^-)$ is the restriction of $E_n(\Sigma^-)$ on Σ^-. Clearly $E_n(\Sigma_1^-)$ is also an inverse closed componentwise independent decomposing algebra.

Inductively we assume that v_1 admits a Riemann–Hilbert factorization relative to Σ_1,

$$v_1 = m_{1-} \theta_1 m_{1+}$$

where

$$\theta_1 = \operatorname{diag}\left[\left(\frac{z - z_+}{z - z_-}\right)^{j_1}, \cdots, \left(\frac{z - z_+}{z - z_-}\right)^{j_n}\right]$$

and $\mathbf{m}_1 \in G\mathcal{E}_n(\Sigma_1) \subset G\mathcal{E}_n(\Sigma)$.

Define on Σ_2, $v_3 = \mathbf{m}_1 v_2 \mathbf{m}_1^{-1} \hat{\theta}_1 \in GE_n(\Sigma_2^-)$ where

$$\hat{\theta}_1 = \operatorname{diag}\left[\left(\frac{z - z_-}{z - z'_-}\right)^{j_1}, \cdots, \left(\frac{z - z_-}{z - z'_-}\right)^{j_n}\right].$$

Inductively we assume that v_3 admits a Riemann–Hilbert factorization relative to Σ_2,

$$v_3 = m_{2-} \theta_2 m_{2+}$$

where

$$\theta_2 = \operatorname{diag}\left[\left(\frac{z - z_+}{z - z'_-}\right)^{k_1}, \cdots, \left(\frac{z - z_+}{z - z'_-}\right)^{k_n}\right]$$

and $\mathbf{m}_2 \in G\mathcal{E}_n(\Sigma_2) \subset G\mathcal{E}_n(\Sigma)$. Set

$$\theta = \operatorname{diag}\left[\left(\frac{z - z_+}{z - z_-}\right)^{k_1}, \cdots, \left(\frac{z - z_+}{z - z_-}\right)^{k_n}\right] \tag{9.2}$$

and $\hat{\theta}_2 = \theta_2 \theta^{-1}$.

Define the componentwise holomorphic function $\mathbf{m}$ on $\mathbb{C} \setminus \Sigma$,

$$\mathbf{m} = \begin{cases} \hat{\theta}_2 \mathbf{m}_2 \hat{\theta}_1^{-1} \mathbf{m}_1 & \text{on } \Omega^+ \\ \mathbf{m}_1 \theta_1 \hat{\theta}_1 \mathbf{m}_2^{-1} \theta_2^{-1} & \text{on } \Omega_1^- \\ \mathbf{m}_1^{-1} \mathbf{m}_2 & \text{on } \Omega_2^- \end{cases} \in G\mathcal{E}_n(\Sigma).$$

Then on Σ_1,

$$\begin{aligned} m_-\theta m_+ &= m_{1-}^{-1}\theta_1\hat{\theta}_1\mathbf{m}_2^{-1}\theta_2^{-1}\theta\hat{\theta}_2\hat{\theta}_1^{-1}m_{1+} \\ &= m_{1-}^{-1}\theta_1 m_{1+} = v_1, \end{aligned}$$

and on Σ_2,

$$\begin{aligned} m_-\theta m_+ &= \mathbf{m}_1^{-1}m_{2-}\theta\hat{\theta}_2 m_{2+}\hat{\theta}_1^{-1}\mathbf{m}_1 \\ &= \mathbf{m}_1^{-1}v_3\hat{\theta}_1^{-1}\mathbf{m}_1 = v_2. \end{aligned}$$

We have the parallel result for $v \in GE_n(\Sigma^+)$.

For general v we write $v = b^-b^+$ such that $b^\pm \in GE_n(\Sigma^-)$. We have already obtained that b^- admits a Riemann–Hilbert factorization

$$b^- = m_-\theta m_+.$$

Since $m_+ - m_+(\infty) \in \operatorname{ran} C_+$, $m_+ \in GE_n(\Sigma^+)$. Therefore $\theta m_+ b^+ \in GE_n(\Sigma^+)$. Thus we have the Riemann–Hilbert factorization

$$\theta m_+ b^+ = m'_-\theta' m'_+.$$

Finally we obtain the Riemann–Hilbert factorization for v,

$$v = (m_- m'_-)\theta' m'_+.$$

The proof for the uniqueness of the partial indices is the same as that in [7].

For the remaining part of the theorem, assume $\det v = 1$. Clearly we may choose m_+ such that $\det m_+(\infty) = 1$. Then the fact that

$$1 = \det v = \det m_- \det\theta \det m_+$$

is a scalar Riemann–Hilbert factorization of 1 forces $\det m_- = \det\theta \det m_+ = 1$. □

In the sequal we replace m_- in Theorem 9.1 by m_-^{-1}. We define the factorized data with respect to a pair of decomposing algebras featured in Theorem 9.1 in the same manner as those previously defined with respect to $\mathbf{H}_I^k(\Sigma^\pm)$.

For a Fredholm operator A we write

$$\alpha(A) = \dim\ker(A); \tag{9.3}$$

$$\beta(A) = \dim\operatorname{coker}(A). \tag{9.4}$$

and the Fredholm index of A as

$$i(A) \overset{\text{def}}{=} \alpha(A) - \beta(A). \tag{9.5}$$

Since $v(\infty) = I$, $m_-(\infty) = m_+(\infty)$. Without loss, we assume $m_\pm(\infty) = I$ (otherwise we may change the basis of $\mathbf{M}_n$). In the following, we show that the classical results regarding the relation between the partial indices and the Fredholm index of the corresponding operators are still valid for the Riemann–Hilbert factorization in a pair of decomposing algebras.

THEOREM 9.2. *Let $k_1, \cdots, k_n$ be the partial indices of v. For any factorized data w of v, we have*

$$\alpha(Id - C_w) = n\sum_{k_j>0} k_j, \tag{9.6}$$

$$\beta(Id - C_w) = -n\sum_{k_j<0} k_j, \tag{9.7}$$

$$i(Id - C_w) = n\sum k_j = \frac{n}{2\pi}\int_\Sigma d\arg\det v. \tag{9.8}$$

Proof. It follows from Proposition 3.4 that we only need to work for a particular factorized data. Here we choose w with $b^+ \stackrel{\text{def}}{=} \theta m_+$ and $b^- \stackrel{\text{def}}{=} m_-$. Then w is a factorized data of v in (9.1). We also define w_1 from $b^\pm \stackrel{\text{def}}{=} m_\pm$, and w_2 from $b^+ = \theta$, $b^- = I$. Then $Id - C_{w_1}$ is invertible because w_1 admits a fundamental solution $m_\pm$. For w_2, (9.6), (9.7), and (9.8) hold by the same proof used in [7]. Now the theorem follows from the relation

$$Id - C_w = (Id - C_{w_1})(Id - C_{w_2}). \tag{9.9}$$ □

We denote by $\dagger$ the Schwarz reflection for the matrix functions, $f^\dagger(z) = f(\bar{z})^*$, and the reflection of a subset of $\mathbb{C}$ about the real axis $\mathbb{R}$. Now we consider the Riemann–Hilbert problem with the contours invariant under the reflection about the real axis. We assume that Σ contains the real axis $\mathbb{R}$; if it does not, then $\Sigma \cap \mathbb{R}$ must be a finite set we may add $\mathbb{R}$ to the contour.

THEOREM 9.3. *Let v be a data supported on a Schwarz reflection invariant contour. Then v has only zero partial indices if* $\operatorname{Re} v \restriction \mathbb{R} \geq 0$, *and* $v = v^\dagger \restriction \Sigma \setminus \mathbb{R}$.

Proof. Clearly we only need to prove the $\mathbf{L}^2$ invertibility of $Id - C_w$ for an arbitrary factorized data w of v. Applying the Schwarz reflection on the Riemann–Hilbert factorization in 9.1 gives

$$v^\dagger = m_+^\dagger \theta^\dagger m_-^{-1\dagger}.$$

It follows from the facts that $\overline{z^\mp} \in \Omega^\mp$ and that

$$(m_\pm^\dagger)^{-1} \in \ker C_\pm$$

that $-k_1, \cdots, -k_n$ are the partial indices of $v^\dagger$. Therefore

$$\beta(Id - C_w) = \alpha(Id - C_{w'}),$$

where w' is a factorized data for $v^\dagger$. Since $v^\dagger$ also satisfies the conditions for v in this theorem, it suffices to show that $\alpha(Id - C_w) = 0$.

Let Ω_ν be a component of $\mathbb{C} \setminus \Sigma$ on the upper half-plane. Then clearly $\Omega_\nu^\dagger$ is a component of $\mathbb{C} \setminus \Sigma$ on the lower half-plane and $\partial\Omega_\nu = (\partial\Omega_\nu)^\dagger$ with the orientation preserved. Let $m \in \ker(Id - C_w)$, then $m_\pm \stackrel{\text{def}}{=} m b^\pm$ give rise to a vanishing solution of the Riemann–Hilbert problem of v. Let $m_{\nu 1}, m_{\nu 2}$ be the boundary values of $\mathbf{m} \restriction \Omega_\nu$, $\mathbf{m} \restriction \Omega_\nu^\dagger$, respectively. Then by the rational approximation (we are working in $\mathbf{L}^2$ space) and a contour integral argument we have

$$\int_{\partial\Omega_\nu} m_{\nu 1} m_{\nu 2}^\dagger = 0.$$

We add them up for all Ω_ν on the upper half-plane to have

$$\sum \int_{\pm\partial\Omega_\nu} m_{\nu 1} m_{\nu 2}^\dagger = 0$$

where $\pm$ is chosen to make $\pm\partial\Omega_\nu$ a positively oriented contour for Ω_ν. Now we assert that

$$\sum \int_{\pm\partial\Omega_\nu} m_{\nu 1} m_{\nu 2}^\dagger = \int_{-\infty}^{\infty} m_- V m_-^\dagger \tag{9.10}$$

where $V = v$ or $v^\dagger$ on different locations. To prove the assertion let Σ_ν be a component of $\Sigma \setminus S$ on the open upper half-plane. Then Σ_ν is part of the common boundary of

a component Ω_ν^+ and a component Ω_ν^- on the upper half-plane. Clearly $\Omega_\nu^{+\dagger}, \Omega_\nu^{-\dagger}$ are components of Ω^-, Ω^+ respectively on the lower half-plane. Now set m_1^+, m_1^- to be the boundary values of $\mathbf{m}$ on Σ_ν from $\Omega_\nu^+, \Omega_\nu^-$, respectively and m_2^+, m_2^- the boundry values of $\mathbf{m}$ on $\Sigma_\nu^\dagger$ from $\Omega_\nu^{-\dagger}, \Omega_\nu^{+\dagger}$, respectively. Clearly the integrals on Σ_ν appear twice in the sum once as

$$\int_{\Sigma_\nu} m_{1+} m_{2-}^\dagger = \int_{\Sigma_\nu} m_{1-} v m_{2-}^\dagger \tag{9.11}$$

and once as

$$\int_{-\Sigma_\nu} m_{1-} m_{2+}^\dagger = \int_{-\Sigma_\nu} m_{1-} v^\dagger m_{2-}^\dagger. \tag{9.12}$$

Since they are off the real axis $v = v^\dagger$, the two integrals cancel. Therefore only the integrals on the real axis survive in the sum. It is easily seen that these integrals are all oriented from left to right and the sum of them gives the right-hand side of (9.10). The assertion is proven. We have

$$\int_{-\infty}^{\infty} m_- V m_-^\dagger = 0. \tag{9.13}$$

Adding this equation to its Schwarz reflection (or Hermitian conjugation) gives

$$\int_{-\infty}^{\infty} m_- (v + v^\dagger) m_-^\dagger = 0. \tag{9.14}$$

Since on $\mathbb{R}$, $\operatorname{Re} v \geq 0$, and $v(\infty) = I$, we have $m_-(z) = 0$ almost everywhere near $z = \infty$ on $\mathbb{R}$. It can be shown by a Cauchy integral argument that the meromorphic extension of m_- near $z = \infty$ equals zero. Since v is nonsingular $m_\pm = 0$ on whole contour Σ. We conclude $\ker(Id - C_w) = 0$. □

For the inverse scattering problem, if v is Hermitian, then $v(z) > 0$ because v is nonsingular and continuous with $v(\infty) = I$. It is shown in [4] that the Riemann–Hilbert problem for the inverse scattering of the $n \times n$ AKNS system with $J + J^* = 0$, and $q + q^* = 0$ is always solvable. Theorem 9.2 implies the following generalization.

If $J + J^* = q + q^* = 0$ for all real z, by the Schwarz reflection principle, $J + J^\dagger = q + q^\dagger = 0$. In this case Σ obviously contains $\mathbb{R}$. Also since the solution $\mathbf{m}$ of the system satisfies $\mathbf{m}^{-1} = \mathbf{m}^\dagger$, it is deduced from this that $v = v^\dagger$ and $\operatorname{Re} v > 0$ on $\mathbb{R}$. Therefore for this kind of system the inverse problem is always solvable.

Finally, for the problem without poles, it follows from the uniqueness of the generalized triangular factorization that $v = v^\dagger$ implies $\Delta = \Delta^\dagger$ in §7. Therefore the conditions for the decay of the potentials at $x = +\infty$ can be explicitly obtained. This also applies to the systems for the generalized wave equation and the generalized sine-Gordon equation for the decay in the nonoblique directions.

Acknowledgment. Professor Adrian Nachman has given many useful suggestions for this work.

REFERENCES

[1] M. J. ABLOWITZ, R. BEALS, AND K. TENENBLAT, *On the solution of the generalized sine-Gordon equations,* Stud. Appl. Math., 74 (1986), pp. 177–203.

[2] M. J. ABLOWITZ, D. J. KAUP, A. C. NEWELL, AND H. SEGUR, *Method for solving the sine-Gordon equation,* Stud. Appl. Math., 53 (1974) pp. 249–315.

[3] R. BEALS AND R. R. COIFMAN, *Scattering and inverse scattering for first order systems,* Comm. Pure Appl. Math., 37 (1984), pp. 39–90.

[4] ———, *Scattering and inverse scattering for first-order systems:* II, Inverse Problems, 3 (1987), pp. 577–593.

[5] B. BOJARSKI, *On the generalized Hilbert boundary–value problem,* Soobshch. Akad. Nauk. Gruzin. SSR, 25 (1960), pp. 385–390. (In Russian.)

[6] P. J. CAUDREY, *The inverse problem for a general $n \times n$ spectral equation,* Phys. D, 6 (1982), pp. 51–66.

[7] K. CLANCY AND I. GOHBERG, *Factorizations of Matrix Functions and Singular Integral Operators,* Birkhäuser Verlag, Basel, Switzerland, 1981.

[8] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations on a half line with kernels depending on the difference of arguments,* Amer. Math. Soc. Transl., 2 (1960), pp. 217– 287.

[9] J-H. LEE, *Analytic properties of Zakharov-Shabat inverse scattering problem with a polynomial spectral dependence of degree n in the potential,* Ph.D. thesis, Yale University, New Haven.

[10] S. PRÖSSDORF, *Some Classes of Singular Equations,* North–Holland, Amsterdam, New York, Oxford, 1978.

[11] L. A. TAKHTAJAN, *Integration of the continuous Heisenberg spin chain through the inverse scattering method,* Phys. Lett. A, 64 (1977), p. 235.

[12] V. E. ZAKHAROV AND A. B. SHABAT, *Integration of nonlinear equations of mathematical physics by the method of inverse scattering,* II, Functional Anal. Appl., 13 (1980), pp. 166–174.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 987–1005, July 1989

013

ASYMPTOTIC EXPANSION OF THE PEARCEY INTEGRAL NEAR THE CAUSTIC*

DAVID KAMINSKI†

Abstract. An asymptotic expansion of the function

$$P(x,y)=\int_{\mathbf{R}}\exp\{i(t^4/4+xt^2/2+yt)\}dt$$

is constructed which remains uniformly valid as $x\to-\infty$ for (x,y) near the caustic $4x^3+27y^2=0$ in the real xy-plane. The result remains valid for a range of complex x and y when P is extended to $\mathbf{C}\times\mathbf{C}$ by analytic continuation.

Key words. uniform asymptotic expansions, Pearcey integral, diffraction integrals, caustics

AMS(MOS) subject classifications. 41A60, 30E15, 33A70

1. Introduction.

The Pearcey function of two real variables x and y is defined by

$$P(x,y)=\int_{-\infty}^{+\infty}\exp\left\{i\left(\frac{t^4}{4}+x\frac{t^2}{2}+yt\right)\right\}dt. \tag{1.1}$$

It is one of a class of "generalized Airy functions"

$$Y_0(x_1,\cdots,x_{m-2})=\int_{-\infty}^{+\infty}\exp\left\{i\left(x_1\theta+\cdots+x_{m-2}\theta^{m-2}+\frac{\theta^m}{m}\right)\right\}d\theta \tag{1.2}$$

(see [7, p. 456]), the usual Airy integral being a constant multiple of $Y_0(x_1)$, and the Pearcey function being $Y_0(y,x/2)$. The role of the oscillatory integrals (1.2) in geometrical optics is well known, $|Y_0|^2$ being related to the intensity of light at caustics characterized by catastrophes which the phase functions represent (see [6, p. 337]).

The function in (1.1) arose in Pearcey's investigation of electromagnetic fields near a "cusp" [11], and so bears his name. More recent work involving (1.1) includes optics [13], scattering theory [5], quantum mechanics [9, p. 172], and the theory of nonlinear waves [8].

Also, just as the Airy function plays an important role in the theory of uniform asymptotic expansions of integrals with two coalescing saddle points [3], the functions Y_0 in (1.2) play a corresponding role in the uniform asymptotic theory of integrals with $m-1$ coalescing saddle points [14]. Consequently, the functions Y_0, for physical as well as mathematical reasons, represent an important class of special functions paralleling the role played by the classical Airy function. We shall restrict our attention to $P(x,y)$ (or, $Y_0(y,x/2)$) in the remainder of this paper.

If we apply the method of stationary phase to (1.1), we find that the asymptotic behaviour of P depends on one or three stationary points, provided one of the variables stays fixed. If we denote the phase function in (1.1) by $\phi(t;x,y)$ and let $\delta=4x^3+27y^2$, then $\phi_t=0$ gives three stationary points if $\delta<0$, and only one if $\delta>0$. However,

* Received by the editors December 9, 1986; accepted for publication March 14, 1987. This work was funded by a National Sciences and Engineering Research Council of Canada Postgraduate Scholarship while the author was at the University of Manitoba.

† Weapons Effectiveness Group, Defence Research Establishment Valcartier, Box 8800, Courcelette, Québec, Canada G0A 1R0.

as $\delta \to 0$, we have some (or all) stationary points coalescing, and in the transition from negative to positive δ, it is not clear how P behaves. Therefore, it is the large negative x-behaviour of (1.1) for (x, y) near the "caustic" $\delta = 0$ that concerns us (if x is positive, then δ is always positive).

If $x \to -\infty$, then for δ to remain small, we must have $|y| \to \infty$. Since $P(x, y)$ is clearly an even function of y, we may restrict ourselves to the case $y > 0$; furthermore, we shall find it to be notationally convenient to replace x by $-x$ in (1.1) and examine $x \to +\infty$. With these conventions, $\delta = 0$ implies $y = 2x^{3/2}/\sqrt{27}$. This suggests that we examine $P(-x, y)$ with $y = \mu x^{3/2}$; when $\mu = 2/\sqrt{27}$, we are right at the caustic. With this relation between x and y, we find that as x and y tend to infinity along $y = \mu x^{3/2}$, μ fixed,

$$P(-x, \mu x^{3/2}) \sim \sum_{j=1}^{3} e^{ix^2 f(t_j;\mu) + \frac{\pi}{4} i \cdot \operatorname{sign}(3t_j^2 - 1)} \sqrt{\frac{2\pi}{x|3t_j^2 - 1|}}. \tag{1.3}$$

Here, $0 \le \mu < 2/\sqrt{27}$, $f(t; \mu) = t^4/4 - t^2/2 + \mu t$, and the t_j are the real roots of $f_t = 0$ (cf. §2 for a description of the t_j). At the caustic $\mu = 2/\sqrt{27}$, we find that

$$P\left(-x, \frac{2x^{3/2}}{\sqrt{27}}\right) \sim \sqrt{\frac{2\pi}{3x}} e^{i(\pi/4 - 2x^2/3)} + \frac{\Gamma(1/3)}{3^{1/3} x^{1/6}} e^{ix^2/12} + \frac{\Gamma(2/3)}{i \cdot 2 \cdot 3^{2/3} \cdot x^{5/6}} e^{ix^2/12} \tag{1.4}$$

and for $\mu > 2/\sqrt{27}$,

$$P(-x, \mu x^{3/2}) \sim \sqrt{\frac{2\pi}{x(3t_0^2 - 1)}} e^{\pi i/4 + ix^2 f(t_0;\mu)} \tag{1.5}$$

where t_0 denotes the only real zero of f_t.

Note that the approximations in (1.3)–(1.5) are of radically differing characters. As $\mu \to (2/\sqrt{27})^-$, t_2 and $t_3 \to 1/\sqrt{3}$ and hence the approximation (1.3) becomes singular; see §2. The important observation to make, at this point, is that as $\delta \to 0$ (i.e., $\mu \to 2/\sqrt{27}$) only two stationary points are coalescing. To exploit this phenomenon, we retain μ as a uniformity parameter and rewrite (1.1) as a sum of two contour integrals, one of which has exactly two relevant, coalescing, saddle points. This allows us to apply a cubic transformation introduced by Chester, Friedman, and Ursell [3], and to construct a uniform asymptotic expansion of (1.1) as $x \to -\infty$ with δ varying in an interval containing 0 (i.e., μ in an interval containing $2/\sqrt{27}$). The expansion we present is in fact valid for certain complex values of the arguments when $P(x, y)$ is extended to $\mathbf{C} \times \mathbf{C}$ by analytic continuation.

At the time of writing, we became aware of a recent publication by Stamnes and Spjelkavik [12] who also noted that only two stationary points coalesce, but their subsequent derivations of asymptotic expansions of the Pearcey function are purely formal. For a brief discussion of their arguments, see §6.

2. Alternate representation. An application of Jordan's inequality shows that, for real x and y, the path of integration in (1.1) may be rotated onto the contour Γ in the complex t-plane, where Γ is the straight line through the origin making an angle of $\pi/8$ with the positive real axis. With this integral representation we can analytically continue $P(x, y)$ to an entire function in $\mathbf{C} \times \mathbf{C}$.

Let

$$\mu = \frac{2}{\sqrt{27}} - \alpha \tag{2.1}$$

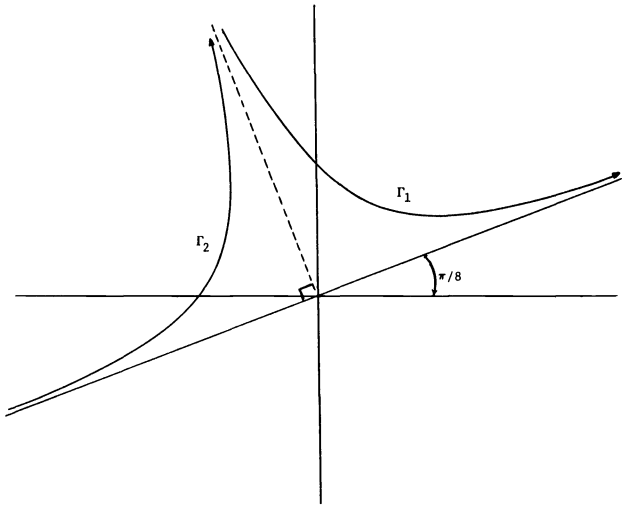


FIG. 1. *The contours* Γ, Γ_1, *and* Γ_2.

and set $y = \mu x^{3/2}$. With x real and positive, the change of variable $t \to x^{1/2}t$ gives

$$P(-x, \mu x^{3/2}) = x^{1/2} \int_{\Gamma} e^{ix^2(t^4/4 - t^2/2 + \mu t)} dt. \tag{2.2}$$

Since this integral converges for all complex μ and for x with $|\arg x| < \pi/4$, subsequent work is valid for these values of x and μ.

By Cauchy's theorem, we may write

$$P(-x, \mu x^{3/2}) = x^{1/2} \int_{\Gamma_1} e^{ix^2(t^4/4 - t^2/2 + \mu t)} dt + x^{1/2} \int_{\Gamma_2} e^{ix^2(t^4/4 - t^2/2 + \mu t)} dt \tag{2.3}$$

where Γ_1 is the contour beginning at $\infty e^{5\pi i/8}$ and ending at $\infty e^{\pi i/8}$, and Γ_2 is the contour beginning at $\infty e^{9\pi i/8}$ and ending at $\infty e^{5\pi i/8}$ (see Fig. 1).

For $i = 1, 2$, set

$$P_i(\lambda; \mu) = \int_{\Gamma_i} e^{i\lambda f(t;\mu)} dt \tag{2.4}$$

where, as in §1, we have put

$$f(t; \mu) = \frac{t^4}{4} - \frac{t^2}{2} + \mu t. \tag{2.5}$$

P_i is analytic for all μ and all λ with $\operatorname{Re} \lambda > 0$.

Since $f_t(t; \mu) = 0$ has the solutions

$$\begin{aligned} t_1(\mu) &= -\frac{2}{\sqrt{3}} \sin\left(\frac{\pi}{3} + \phi\right) \\ t_2(\mu) &= \frac{2}{\sqrt{3}} \sin \phi \\ t_3(\mu) &= \frac{2}{\sqrt{3}} \sin\left(\frac{\pi}{3} - \phi\right) \end{aligned}$$

for $\Delta = -4 + 27\mu^2 < 0$, with ϕ given by

$$3\phi = \arcsin\left(\mu\frac{\sqrt{27}}{2}\right), \ |\phi| \leq \pi/6$$

(see [10]), we see that as $\Delta \to 0^-$ (i.e., $\mu \to (2/\sqrt{27})^-$), $t_1 \to -2/\sqrt{3}$, $t_2 \to 1/\sqrt{3}$, $t_3 \to 1/\sqrt{3}$. Thus the two positive roots of $f_t = 0$ coalesce as $\mu \to 2/\sqrt{27}$, but remain well separated from the negative root.

If $\Delta > 0$, we can invoke Cardan's formulae (or use the trigonometric solution above with ϕ complex) to observe the same phenomenon, only this time, as $\Delta \to 0^+$, the complex conjugate pair of roots of $f_t = 0$ coalesce to $1/\sqrt{3}$ with the real root remaining isolated.

Also, by examining $\sin\phi$ and $\sin(\pi/3 - \phi)$ for $\phi \in]0, \pi/6[$, we find that $t_3(\mu) > 1/\sqrt{3} > t_2(\mu)$ for $\mu < 2/\sqrt{27}$.

With this background, we shall obtain the large λ behaviour of the $P_i(\lambda;\mu)$ as $\lambda \to \infty$. This in turn will yield the desired expansion of (2.3) as $x \to \infty$ uniformly for $\mu \in [\mu_0, \mu_1]$ where $0 < \mu_0 < 2/\sqrt{27} < \mu_1$.

3. Uniform expansion of P_1. Since P_1 has two relevant coalescing saddle points, this suggests invoking a cubic transformation as is done in [3] or [2, §9.2]. For convenience, we shall choose α, given in (2.1), as our uniformity parameter since the saddles coincide when α vanishes.

As the saddles coalesce at $t = 1/\sqrt{3}$, we develop f as follows:

$$\begin{aligned} f(t;\mu) &= -\frac{\alpha}{\sqrt{3}} + \frac{1}{12} - \alpha z + \frac{z^3}{\sqrt{3}} + \frac{z^4}{4} \\ &\equiv g(z;\alpha) - \frac{\alpha}{\sqrt{3}} + \frac{1}{12} \end{aligned}$$

where we have set $z = t - 1/\sqrt{3}$. Thus

$$P_1(\lambda;\mu) = e^{-i\lambda(\alpha/\sqrt{3} - 1/12)} \int_{\Gamma_1'} e^{i\lambda g(z;\alpha)} dz$$

where Γ_1' is the translate of Γ_1 by $1/\sqrt{3}$. Note that the zeroes of $g_z(z;\alpha)$ are given by

$$\begin{aligned} z_1(\alpha) &= -\frac{2}{\sqrt{3}}\sin\left(\phi + \frac{\pi}{3}\right) - \frac{1}{\sqrt{3}} \\ z_2(\alpha) &= \frac{2}{\sqrt{3}}\sin\phi - \frac{1}{\sqrt{3}} \\ z_3(\alpha) &= \frac{2}{\sqrt{3}}\sin\left(\frac{\pi}{3} - \phi\right) - \frac{1}{\sqrt{3}} \end{aligned}$$

where

$$3\phi = \arcsin\left(1 - \frac{\alpha\sqrt{27}}{2}\right), \ |\phi| \leq \frac{\pi}{6}.$$

As in §2, by examining $\sin\phi$ and $\sin(\pi/3 - \phi)$, we have $z_1(\alpha) < -1 < z_2(\alpha) < 0 < z_3(\alpha)$ for $\alpha > 0$ sufficiently small. When $\alpha < 0$, z_2 and z_3 are complex conjugates. In either case, $\alpha \to 0$ implies $z_2(\alpha), z_3(\alpha) \to 0$.

Now we introduce the change of variables

$$g(z;\alpha) = \frac{u^3}{3} - \zeta u + \eta \tag{3.1}$$

where $\zeta = \zeta(\alpha)$ and $\eta = \eta(\alpha)$ are to be determined. In order that (3.1) defines an analytic function $u(z)$ near z_2 and z_3, we require that z_2 and z_3 correspond to $-\zeta^{1/2}$ and $\zeta^{1/2}$ respectively. Accordingly, we find that

$$\zeta^{3/2} = \frac{3}{4}[g(z_2;\alpha) - g(z_3;\alpha)] \tag{3.2}$$

and

$$\eta = \frac{1}{2}[g(z_2;\alpha) + g(z_3;\alpha)]. \tag{3.3}$$

Equation (3.1) may be solved explicitly to give three possible candidates for our change of variables: for $\zeta \neq 0$,

$$\begin{aligned} u_1(z;\alpha) &= -2\zeta^{1/2}\sin(\pi/3 + \psi) \\ u_2(z;\alpha) &= 2\zeta^{1/2}\sin\psi \\ u_3(z;\alpha) &= 2\zeta^{1/2}\sin(\pi/3 - \psi) \end{aligned}$$

with

$$\sin 3\psi = \frac{3(\eta - g(z;\alpha))}{2\zeta^{3/2}};$$

for $\zeta = 0$, we have $u = (3[g(z) - \eta])^{1/3}$ and, again, there are three branches. Please note that we frequently omit mentioning the explicit dependence of g on α.

From (3.2) and (3.3), we have

$$\sin 3\psi = \frac{\frac{3}{2}[g(z_3) - g(z_2)]}{2\zeta^{3/2}} = -1$$

when $z = z_2$, and when $z = z_3$ we get

$$\sin 3\psi = \frac{\frac{3}{2}[g(z_2) - g(z_3)]}{2\zeta^{3/2}} = +1.$$

Thus, we have

$$\begin{aligned} u_1(z_2;\alpha) &= -\zeta^{1/2} & u_1(z_3;\alpha) &= -2\zeta^{1/2} \\ u_2(z_2;\alpha) &= -\zeta^{1/2} & u_2(z_3;\alpha) &= \zeta^{1/2} \\ u_3(z_2;\alpha) &= 2\zeta^{1/2} & u_3(z_3;\alpha) &= \zeta^{1/2}. \end{aligned}$$

Therefore, $u_2(z;\alpha)$ is the desired uniformly analytic solution to (3.1). We now set $u(z;\alpha) \equiv u_2(z;\alpha)$.

Before continuing further, let us briefly examine the nature of the mapping (3.1). From Chester, et al. [3], we know that $u(z;\alpha)$ is uniformly analytic and one-to-one near $z = 0$. Much more, however, can be said. In the following we will show that the mapping (3.1) in the present case is in fact one-to-one and analytic along the contour Γ_1'.

In the subsequent analysis, assume $\alpha > 0$. The mapping from the z- to u-planes can be most easily studied by introducing intermediate variables Z, ψ given by

$$\begin{aligned} Z &= g(z;\alpha) - \eta \\ \sin 3\psi &= -3Z/2\zeta^{3/2} \\ u &= 2\zeta^{1/2}\sin\psi; \end{aligned}$$

see, for example, Copson [4]. We begin by decomposing the z-plane into four regions, each of which maps onto the Z-plane. A useful device in this process is the determination of all curves that g sends to the real Z-axis. We adopt the notational convention of denoting that part of a region Ω that has positive imaginary part by Ω^+.

The mapping $Z = g(z;\alpha) - \eta$ clearly takes the real line in the z-plane to the real line in the Z-plane in the following manner:

$$\begin{aligned} &Z([z_3, +\infty[) = [-\tfrac{2}{3}\zeta^{3/2}, +\infty[\\ &Z([z_2, z_3]) = [-\tfrac{2}{3}\zeta^{3/2}, \tfrac{2}{3}\zeta^{3/2}] \\ &Z([z_1, z_2]) = [g(z_1) - \eta, \tfrac{2}{3}\zeta^{3/2}] \\ &Z(]-\infty, z_1]) = [g(z_1) - \eta, +\infty[, \end{aligned} \tag{3.4}$$

the mapping Z being one-to-one on each interval. The remaining curves are the nontrivial solutions of

$$\begin{aligned} \operatorname{Im}(g(z;\alpha) - \eta) = 0 \quad &= \quad \operatorname{Im} g(z;\alpha) \\ &= \quad \operatorname{Im}(g(z;\alpha) - g(z_k;\alpha)), \ k = 1,2,3 \end{aligned} \tag{3.5}$$

since each critical point z_k being real implies that $g(z_k;\alpha)$ is real. Note that since the left half of (3.5) is independent of k, the solution curves arising from the right half of (3.5) are the same for each k.

Develop g about the critical point $z = z_k$. Then (3.5) becomes

$$\operatorname{Im}\left(\frac{g''(z_k;\alpha)}{2}t^2 + \frac{g'''(z_k;\alpha)}{6}t^3 + \frac{1}{4}t^4\right) = 0$$

where we have set $t = z - z_k = \sigma + i\tau$, and suppressed dependence on k. Thus

$$0 = \frac{2\sqrt{3}z_k + z_k^2}{2}\cdot 2\sigma\tau + \frac{2\sqrt{3} + 6z_k}{6}(3\sigma^2\tau - \tau^3) + (\sigma^3\tau - \sigma\tau^3)$$

or, since we have accounted for $\tau \equiv 0$ in (3.4),

$$0 = (2\sqrt{3}z_k + 3z_k^2)\sigma + \frac{2\sqrt{3} + 6z_k}{2}\sigma^2 + \sigma^3 - \left(\frac{2\sqrt{3} + 6z_k}{6} + \sigma\right)\tau^2$$

yielding

$$\tau = \pm\sqrt{\frac{(2\sqrt{3}z_k + 3z_k^2)\sigma + \frac{2\sqrt{3}+6z_k}{2}\sigma^2 + \sigma^3}{\frac{2\sqrt{3}+6z_k}{6} + \sigma}} \tag{3.6}$$

wherever the quantity inside the outer surd is nonnegative.

Since $\sigma = \tau = 0$ satisfies (3.6) and $\sigma + i\tau = z - z_k$, each z_k lies on at least one of the curves determined by (3.6). Equation (3.6) gives rise to three curves; two of

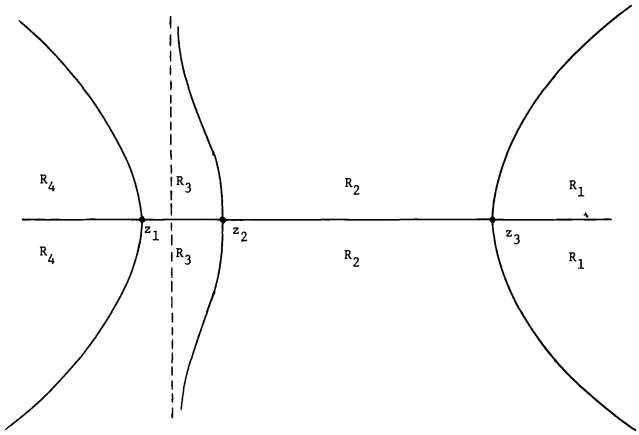


FIG. 2. *Curves that map to the real Z-axis (solid curves) and the regions R_j, $j = 1, 2, 3, 4$.*

these curves have the 45^0 lines through the origin as asymptotes, and the other (the one containing z_2) has the vertical line $\sigma = -(2\sqrt{3} + 6z_k)/6$ as its asymptote. These curves are displayed in Fig. 2.

The curves that do not lie along the real z-axis partition the z-plane into four regions, R_1, R_2, R_3, and R_4, each of which maps in a one-to-one fashion onto the Z-plane (on the boundaries, the map g may be two-to-one) (see Fig. 2).

Consider R_1^+ and the map $z \to g(z;\alpha) - \eta$. From (3.4) we know that $[z_3, +\infty[$ maps onto $[-\frac{2}{3}\zeta^{3/2}, +\infty[$, and the curve bounding the upper extent of R_1^+ is mapped to $]-\infty, -\frac{2}{3}\zeta^{3/2}]$. Thus we see that R_1^+ maps to the upper half of the Z-plane. As well, the steepest descent curve of $ig(z;\alpha)$ from $z = z_3$ to $\infty e^{\pi i/8}$ lies within R_1^+ and maps to a vertical line in the Z-plane running through $Z = -\frac{2}{3}\zeta^{3/2}$. In Figs. 3-5, shaded regions map to shaded regions, dotted lines represent steepest descent curves, and in each figure, letters that correspond represent points that correspond under each indicated map (see Fig. 3).

Under the map $\sin 3\psi = -3Z/2\zeta^{3/2}$, we note the following: $Z = -\frac{2}{3}\zeta^{3/2}$ is mapped to $\psi = \pi/6$; with $\psi \in [\pi/6, \pi/2]$, Z is real and in $[-\frac{2}{3}\zeta^{3/2}, \frac{2}{3}\zeta^{3/2}]$, and the rays $\pi/6 + i\tau$, $\pi/2 + i\tau$, $\tau \geq 0$ map to the segments $]-\infty, -\frac{2}{3}\zeta^{3/2}]$ and $[\frac{2}{3}\zeta^{3/2}, +\infty[$ in the Z-plane, respectively. Moreover, a straightforward calculation reveals that the steepest descent curve lying in the Z-plane maps to the curve $\gamma = \frac{1}{3} \cdot \log[(1 - \cos 3\beta)/(\sin 3\beta)]$, $\psi = \beta + i\gamma$, $\gamma \geq 0$, $\beta \leq \pi/3$, in the ψ-plane (see Fig. 3).

The map $u = 2\zeta^{1/2} \sin\psi$ takes the interval $[\frac{\pi}{6}, \frac{\pi}{2}]$ in the ψ-plane to $[\zeta^{1/2}, 2\zeta^{1/2}]$ in the u-plane, and maps $\frac{\pi}{2} + i\tau$, $\tau \geq 0$ onto $[2\zeta^{1/2}, +\infty[$. The image of the ψ-plane ray $\frac{\pi}{6} + i\tau$, $\tau \geq 0$, is the curve

$$u = \zeta^{1/2}[\cosh\tau + i\sqrt{3}\sinh\tau],$$

which we see as being the first quadrant branch of the hyperbola $(\operatorname{Re} u)^2 - \frac{1}{3}(\operatorname{Im} u)^2 = \zeta$ (to see this, put $\operatorname{Re} u = \zeta^{1/2}\cosh\tau$, $\operatorname{Im} u = \sqrt{3}\zeta^{1/2}\sinh\tau$ and employ the identity $\cosh^2\tau - \sinh^2\tau = 1$). The steepest descent curve in the ψ-plane maps to the steepest descent curve of $i(u^3/3 - \zeta u + \eta)$ beginning at $u = \zeta^{1/2}$ and ending at $\infty e^{\pi i/6}$. This is most easily seen by examining $Z = u^3/3 - \zeta u + \eta$ directly; see Fig. 3.

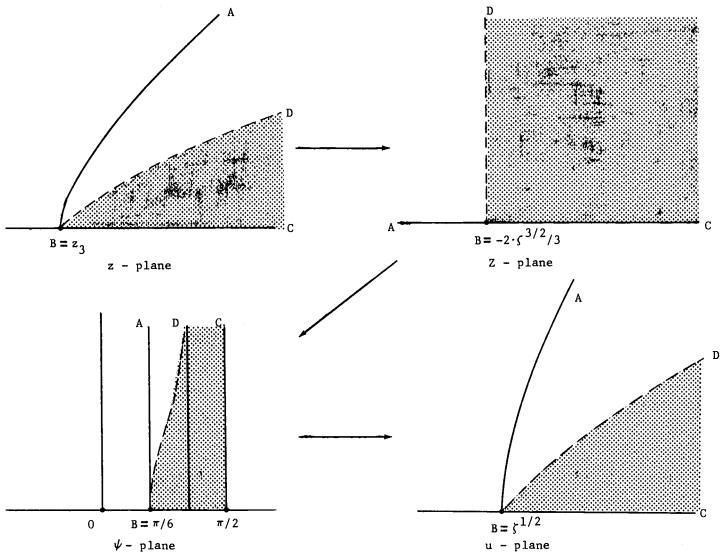


FIG. 3. *Effect of the mappings* $Z = g(z;\alpha) - \eta$, $Z = -\frac{2}{3}\zeta^{3/2} \cdot \sin 3\psi$, $u = 2\zeta^{1/2} \sin\psi$ *on* R_1^+.

In a completely similar fashion, we find the images of R_2^+ and R_3^+ under the sequence of transformations $z \to Z \to \psi \to u$. The effect of these maps are presented in Figs. 4-5.

Collecting these maps together gives a conformal map from $R_1^+ \cup R_2^+ \cup R_3^+$ to the upper half of the u-plane which is one-to-one, although the intermediate transformations fail to be one-to-one when applied to all of $R_1^+ \cup R_2^+ \cup R_3^+$.

In the event that $\alpha < 0$, we would proceed as before, this time using (3.5) with $k = 1$ since (3.6) holds only for the real saddle z_1. Note that Im $\eta = 0$ for $\alpha < 0$.

If we take Γ_1' to be the steepest descent curve beginning at $\infty e^{5\pi i/8}$ and ending at $z = z_2$, followed by the straight line segment $[z_2, z_3]$, and thereafter the steepest descent curve from $z = z_3$ to $\infty e^{\pi i/8}$, then $u(z;\alpha)$ maps Γ_1' onto the curve formed by steepest descent curves from $-\zeta^{1/2}$ to $\infty e^{5\pi i/6}$, and $+\zeta^{1/2}$ to $\infty e^{\pi i/6}$, along with the line segment $[-\zeta^{1/2}, \zeta^{1/2}]$, suitably oriented. Call the image of Γ_1' in the u-plane C. We may now write

$$e^{i\lambda(\alpha/\sqrt{3}-1/12)} P_1(\lambda;\mu) = \int_C e^{i\lambda(u^3/3-\zeta u+\eta)} g_0(u;\alpha)\,du$$

where we have put $g_0(u;\alpha) = dz/du$. Define, as in [2], the function sequences $\{g_n\}$, $\{h_n\}$, $\{p_n\}$, and $\{q_n\}$ by

$$\begin{aligned} p_n(\alpha) &= \tfrac{1}{2}[g_n(\zeta^{1/2};\alpha) + g_n(-\zeta^{1/2};\alpha)] \\ q_n(\alpha) &= \tfrac{1}{2\zeta^{1/2}}[g_n(\zeta^{1/2};\alpha) - g_n(-\zeta^{1/2};\alpha)] \\ g_n(u;\alpha) &= p_n(\alpha) + q_n(\alpha)u + (u^2 - \zeta)h_n(u;\alpha) \\ g_{n+1}(u;\alpha) &= \tfrac{d}{du}h_n(u;\alpha), \quad n = 0, 1, 2, \cdots. \end{aligned} \tag{3.7}$$

Then, by successive substitution and partial integration, we obtain

$$e^{i\lambda(\alpha/\sqrt{3}-1/12)} P_1(\lambda;\mu) \sim e^{i\lambda\eta} \sum_{n=0}^{\infty} \left(\frac{i}{\lambda}\right)^n [p_n(\alpha)F(\lambda;\zeta) + q_n(\alpha)G(\lambda;\zeta)]$$

as $\lambda \to \infty$, uniformly in ζ. Here,

$$F(\lambda;\zeta) = \int_C e^{i\lambda(u^3/3-\zeta u)}\,du$$

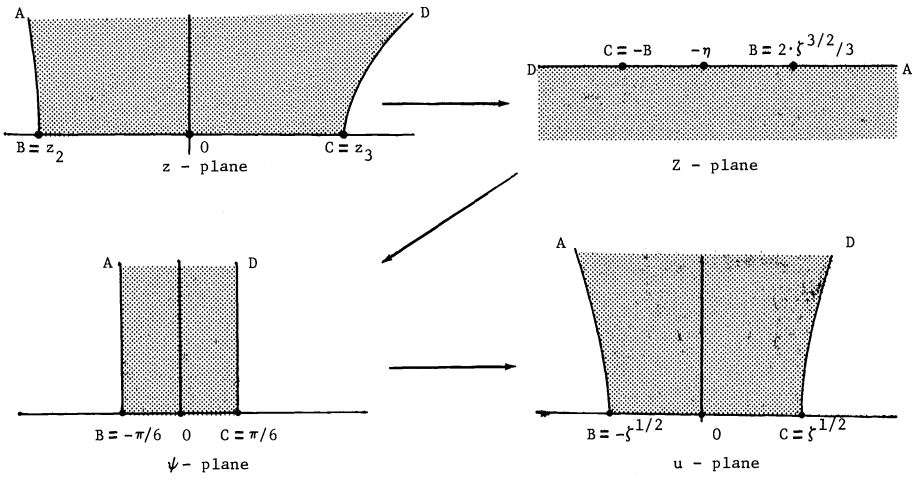


FIG. 4. *Effect of the sequence of maps* $z \to Z \to \psi \to u$ *applied to* R_2^+.

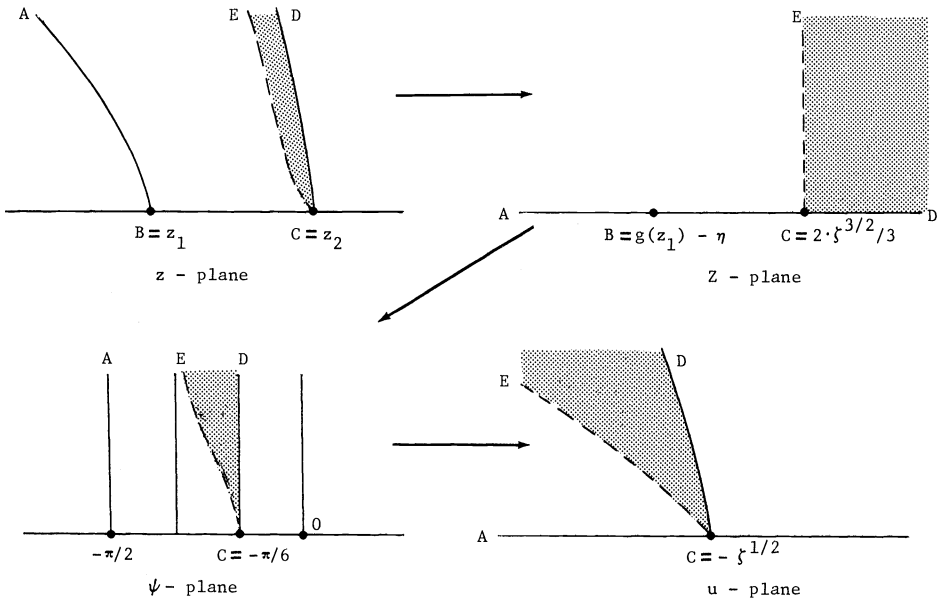


FIG. 5. *Effect of the sequence of maps* $z \to Z \to \psi \to u$ *applied to* R_3^+.

$$G(\lambda;\zeta) = \int_C ue^{i\lambda(u^3/3-\zeta u)}du,$$

both integrals converging absolutely and uniformly for $\operatorname{Re}\lambda > 0$, $\zeta \in \mathbf{C}$. By a simple change of variables, we see that in $\operatorname{Re}\lambda > 0$,

$$F(\lambda;\zeta) = \frac{2\pi}{\lambda^{1/3}} \cdot Ai(-\lambda^{2/3}\zeta)$$
$$G(\lambda;\zeta) = \frac{2\pi}{i\lambda^{2/3}} \cdot Ai'(-\lambda^{2/3}\zeta).$$

Hence,

$$e^{i\lambda(\alpha/\sqrt{3}-1/12)}P_1\left(\lambda;\frac{2}{\sqrt{27}}-\alpha\right) \sim e^{i\lambda\eta}\cdot\frac{2\pi}{\lambda^{1/3}} \cdot\sum_{n=0}^{\infty}\left[p_n(\alpha)Ai(-\lambda^{2/3}\zeta)+\frac{1}{i\lambda^{1/3}}q_n(\alpha)Ai'(-\lambda^{2/3}\zeta)\right]\left(\frac{i}{\lambda}\right)^n$$

as $\lambda \to \infty$, uniformly in α near zero. In §6, we shall have occasion to use the approximation

$$P_1\left(\lambda;\frac{2}{\sqrt{27}}-\alpha\right) = e^{i\lambda(1/12-\alpha/\sqrt{3}+\eta)} \cdot\left\{p_0(\alpha)\frac{2\pi}{\lambda^{1/3}}\cdot Ai(-\lambda^{2/3}\zeta)\cdot\left[1+O\left(\frac{1}{\lambda}\right)\right] + q_0(\alpha)\frac{2\pi}{i\lambda^{2/3}}\cdot Ai'(-\lambda^{2/3}\zeta)\left[1+O\left(\frac{1}{\lambda}\right)\right]\right\} \tag{3.8}$$

for $\lambda \to \infty$ as above.

4. The coefficients $p_0(\alpha)$ and $q_0(\alpha)$. Although expressions for $p_0(\alpha)$ and $q_0(\alpha)$ can be calculated routinely, because of the labour involved in obtaining limiting forms as $\alpha \to 0$, we reproduce some of the requisite analysis. The reader will appreciate the difficulty in calculating higher coefficients from these two examples.

We begin by remarking that the following analysis remains valid for all (small) complex α, provided we choose the principal branch for $\alpha^{1/2}$. For the purpose of exposition, we shall take $\alpha > 0$.

Differentiate (3.1) twice with respect to u; this gives

$$g_{zz}\left(\frac{dz}{du}\right)^2 + g_z\left(\frac{d^2z}{du^2}\right) = 2u.$$

Evaluation at $z = z_3$, and the use of the fact that z_3 corresponds with $\zeta^{1/2}$, gives

$$(2\sqrt{3}z_3 + 3z_3^2)\left(\frac{dz}{du}\right)^2\Big|_{u=\zeta^{1/2}} = 2\zeta^{1/2}$$

whence

$$\frac{dz}{du}\Big|_{u=\zeta^{1/2}} = \pm\sqrt{\frac{2\zeta^{1/2}}{2\sqrt{3}z_3+3z_3^2}}.$$

Since z_3, for $\alpha > 0$, is a local minimum of $g(z;\alpha)$, $g_{zz}(z_3;\alpha)$ is positive, and so the expression under the square root is positive. Also, as z increases with u,

$$\frac{dz}{du}\Big|_{u=\zeta^{1/2}} = \sqrt{\frac{2\zeta^{1/2}}{2\sqrt{3}z_3 + 3z_3^2}}.$$

Similarly,

$$\frac{dz}{du}\Big|_{u=-\zeta^{1/2}} = \sqrt{\frac{-2\zeta^{1/2}}{2\sqrt{3}z_2 + 3z_2^2}}.$$

Since $g_0(u;\alpha) = dz/du$, we have, from the first two equations of (3.7) with $n = 0$,

$$p_0(\alpha) = \frac{1}{2}\left\{\sqrt{\frac{2\zeta^{1/2}}{2\sqrt{3}z_3 + 3z_3^2}} + \sqrt{\frac{-2\zeta^{1/2}}{2\sqrt{3}z_2 + 3z_2^2}}\right\} \tag{4.1}$$

$$q_0(\alpha) = \frac{1}{2\zeta^{1/2}}\left\{\sqrt{\frac{2\zeta^{1/2}}{2\sqrt{3}z_3 + 3z_3^2}} - \sqrt{\frac{-2\zeta^{1/2}}{2\sqrt{3}z_2 + 3z_2^2}}\right\}. \tag{4.2}$$

Use of equation (3.2) allows us to express (4.1) and (4.2) completely in terms of the function g.

$p_0(\alpha)$ and $q_0(\alpha)$ can also be obtained from relations involving only z_1, the zero of g' that remains isolated from z_2 and z_3. To see this, we observe

$$\begin{aligned} g'(z) &= z^3 + \sqrt{3}z^2 - \alpha \\ &= z^3 - (z_1 + z_2 + z_3)z^2 + (z_1z_2 + z_1z_3 + z_2z_3)z - z_1z_2z_3, \end{aligned}$$

and hence

$$\begin{aligned} z_1 + z_2 + z_3 &= -\sqrt{3}, \\ z_1z_2 + z_1z_3 + z_2z_3 &= 0, \\ z_1z_2z_3 &= \alpha. \end{aligned}$$

Use of the latter three equations permits the expression of ζ^3, p_0q_0, and p_0/q_0, in terms of α and z_1. For example, we find

$$p_0q_0 = \frac{z_1^2}{2\alpha}\left[\frac{3z_1^2 + 2\sqrt{3}z_1 - 3}{6\sqrt{3}z_1^2 + 6z_1 - 9\alpha}\right].$$

However, these expressions are not appreciably simpler than (4.1) and (4.2), in view of the symmetry we exploit below.

In the limit $\alpha \to 0^+$, we find that $\phi \to \frac{\pi}{6}^-$. Write

$$\phi = \frac{\pi}{6} - \theta.$$

Then

$$z_2 = \frac{1}{\sqrt{3}}[\cos(-\theta) + \sqrt{3}\sin(-\theta) - 1]$$

and

$$z_3 = \frac{1}{\sqrt{3}}[\cos(\theta) + \sqrt{3}\sin(\theta) - 1].$$

If we let

$$\chi(\theta) = \frac{1}{\sqrt{3}}[\cos(\theta) + \sqrt{3}\sin(\theta) - 1], \tag{4.3}$$

we have $z_2 = \chi(-\theta), z_3 = \chi(\theta)$. Furthermore, from $\sin 3\phi = \arcsin(1 - \sqrt{27}\alpha/2)$, we have

$$\phi = \frac{\pi}{6} - \frac{1}{3} \cdot 3^{3/4}\alpha^{1/2}\left(1 + \frac{\sqrt{3}}{8}\alpha + \frac{81\alpha^2}{32 \cdot 20} + O(\alpha^3)\right)$$

so that

$$\theta = \frac{\alpha^{1/2}}{3^{1/4}}\left[1 + \frac{\sqrt{3}}{8}\alpha + \frac{81\alpha^2}{32 \cdot 20} + O(\alpha^3)\right]. \tag{4.4}$$

The Maclaurin series for sine involves only odd powers, and that for cosine only even powers. Hence, the approximation for $z_3(\alpha)$ gives

$$z_3(\alpha) = \frac{\alpha^{1/2}}{3^{1/4}} - \frac{\alpha}{6} + \frac{5 \cdot 3^{1/4}}{72}\alpha^{3/2} - \frac{3^{1/2}}{27}\alpha^2 + O(\alpha^{5/2}).$$

The corresponding approximation for $z_2(\alpha)$ follows by replacing $\alpha^{1/2}$ by $-\alpha^{1/2}$ in that for $z_3(\alpha)$.

Indeed, any function of z_3 gives rise to a function of z_2 by the process of replacing $\alpha^{1/2}$ by $-\alpha^{1/2}$. Using this, we have

$$g(z_3) = -2 \cdot 3^{-5/4}\alpha^{3/2}\left[1 - \frac{3^{1/4}}{8}\alpha^{1/2} + \frac{3^{1/2}}{24}\alpha + O(\alpha^{3/2})\right]$$

and

$$g(z_2) = 2 \cdot 3^{-5/4}\alpha^{3/2}\left[1 + \frac{3^{1/4}}{8}\alpha^{1/2} + \frac{3^{1/4}}{24}\alpha + O(\alpha^{3/2})\right].$$

These two equations and (3.2) imply

$$\zeta^{1/2}(\alpha) = 3^{-1/12}\alpha^{1/2}\left[1 + \frac{\alpha}{24\sqrt{3}} + O(\alpha^2)\right],$$

where we have made use of the binomial formula.

Continuing, we get

$$g_{zz}(z_3) = 2 \cdot 3^{1/4}\alpha^{1/2}\left[1 + 3^{-3/4}\alpha^{1/2} - \frac{7 \cdot 3^{-1/2}}{24}\alpha + O(\alpha^{3/2})\right]$$

and

$$g_{zz}(z_2) = -2 \cdot 3^{1/4}\alpha^{1/2}\left[1 - 3^{-3/4}\alpha^{1/2} - \frac{7 \cdot 3^{-1/2}}{24}\alpha + O(\alpha^{3/2})\right]$$

whence

$$\sqrt{\frac{-2\zeta^{1/2}}{2\sqrt{3}z_2+3z_2^2}} = 3^{-1/6}\left[1+\frac{3^{-3/4}}{2}\alpha^{1/2}+O(\alpha)\right]$$

and

$$\sqrt{\frac{2\zeta^{1/2}}{2\sqrt{3}z_3+3z_3^2}} = 3^{-1/6}\left[1-\frac{3^{-3/4}}{2}\alpha^{1/2}+O(\alpha)\right].$$

These two approximations (together with that obtained for $\zeta^{1/2}$) in formulae (4.1) and (4.2) yield

$$p_0(\alpha) = 3^{-1/6}[1+O(\alpha)], \tag{4.5}$$

and

$$q_0(\alpha) = -\frac{3^{-5/6}}{2}[1+O(\alpha)], \tag{4.6}$$

as $\alpha \to 0$.

5. Expansion of P_2. Because Γ_2 begins and ends in valleys at ∞, we see that the contour can be deformed into the steepest descent path leading away from the saddle point t_1. The determination of the steepest descent path follows.

We begin by expanding f about the point $t = t_1$:

$$if(t;\mu) = if(t_1;\mu) + i\left(\frac{3t_1^2-1}{2}\right)(t-t_1)^2 + it_1(t-t_1)^3 + \frac{i}{4}(t-t_1)^4.$$

Let $z = t - t_1 = x + iy$. Steepest descent paths are among the level curves $0 = \mathrm{Im}[if(t;\mu) - if(t_1;\mu)]$. Consequently,

$$\frac{1}{2}(3t_1^2-1)(x^2-y^2) + t_1(x^3-3xy^2) + \frac{1}{4}(x^4-6x^2y^2+y^4) = 0$$

gives steepest paths. Solving for y as a function of x gives

$$y = \pm\frac{1}{\sqrt{2}}\left\{6x^2+12t_1x+2(3t_1^2-1) \pm \sqrt{(6x^2+12t_1x+2(3t_1^2-1))^2 - 4(x^4+4t_1x^3+2(3t_1^2-1)x^2)}\right\}^{1/2}$$

wherever the real square roots are defined. We see that the steepest descent curve through t_1 begins at $\infty e^{-7\pi i/8}$ and ends at $\infty e^{5\pi i/8}$. Fig. 6 displays hills and valleys in the case $\delta = 0$ (or $\mu = 2/\sqrt{27}$). Shaded regions represent valleys, the unshaded regions being hills. Solid curves (excluding the real axis) represent steepest descent or ascent paths according as they lie in shaded or unshaded regions respectively.

Let Γ_{2+} be the steepest descent curve running from t_1 to $\infty e^{5\pi i/8}$, and let Γ_{2-} be the steepest descent contour running from t_1 to $\infty e^{-7\pi i/8}$. Then we have

$$P_2(\lambda;\mu) = \int_{\Gamma_{2+}} e^{i\lambda f}dt - \int_{\Gamma_{2-}} e^{i\lambda f}dt.$$

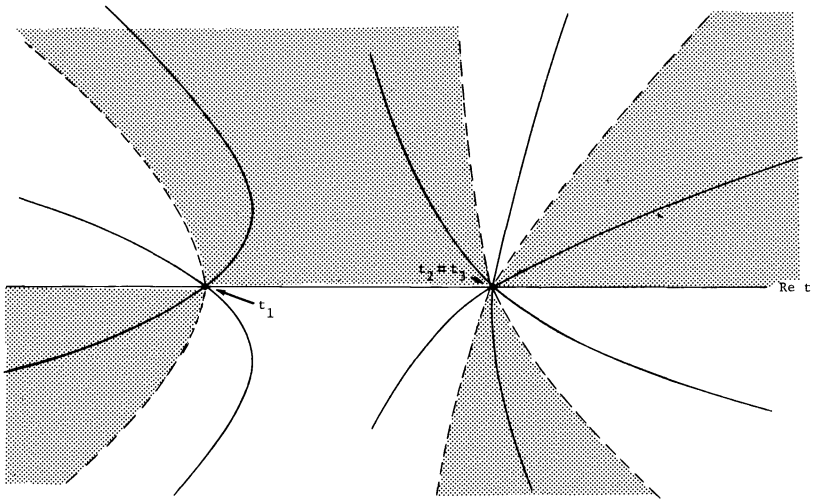


FIG. 6. *Hills and valleys defined by steepest curves through* t_1.

Let $w = if(t_1;\mu) - if(t;\mu)$ so that on $\Gamma_{2\pm}$, w is positive and increasing as we move away from t_1.

Let t_+ be the solution of $w = if(t_1;\mu) - if(t;\mu)$ on Γ_{2+}, t_- being the corresponding solution on Γ_{2-}. By the Lagrange inversion theorem, since

$$w = -\frac{i}{2}(3t_1^2 - 1)(t_\pm - t_1)^2 - it_1(t_\pm - t_1)^3 - \frac{i}{4}(t_\pm - t_1)^4$$

on $\Gamma_{2\pm}$ respectively, there are numbers $a_n^\pm$ for which

$$t_\pm = t_1 + \sum_{n=1}^{\infty} \frac{a_n^\pm}{n!} w^{n/2}$$

whence

$$\frac{dt_+}{dw} - \frac{dt_-}{dw} = \sum_{n=1}^{\infty} \frac{a_n^+ - a_n^-}{2(n-1)!} w^{\frac{n}{2}-1}.$$

Thus

$$P_2(\lambda;\mu) \sim e^{i\lambda f(t_1;\mu)} \sum_{n=1}^{\infty} \frac{a_n^+ - a_n^-}{2(n-1)!} \frac{\Gamma(n/2)}{\lambda^{n/2}}$$

as $\lambda \to \infty$; see arguments in [4, §§30 and 33].

The first few terms are easily found to be

$$a_1^+ = e^{\pi i/4} \left(\frac{2}{3t_1^2 - 1}\right)^{1/2}$$

$$a_2^+ = -i \left(\frac{2}{3t_1^2 - 1}\right) \cdot \left(\frac{2t_1}{3t_1^2 - 1}\right)$$

$$
\begin{aligned}
a_3^+ &= e^{3\pi i/4}\left(\frac{2}{3t_1^2-1}\right)^{3/2}\cdot\left[\frac{-3}{2(3t_1^2-1)}+\frac{15t_1^2}{2(3t_1^2-1)^2}\right]\\
a_1^- &= e^{5\pi i/4}\left(\frac{2}{3t_1^2-1}\right)^{1/2}\\
a_2^- &= a_2^+\\
a_3^- &= e^{-\pi i/4}\left(\frac{2}{3t_1^2-1}\right)^{3/2}\cdot\left[\frac{-3}{2(3t_1^2-1)}+\frac{15t_1^2}{2(3t_1^2-1)^2}\right]
\end{aligned}
$$

whence

$$
P_2(\lambda;\mu)=e^{i\lambda f(t_1;\mu)}\cdot\left\{\sqrt{\frac{\pi}{3t_1^2-1}}\frac{1+i}{\lambda^{1/2}}+\frac{(-3-6t_1^2)\pi^{1/2}}{4\cdot(3t_1^2-1)^{7/2}}\frac{1-i}{\lambda^{3/2}}+O\left(\frac{1}{\lambda^{5/2}}\right)\right\} \tag{5.1}
$$

where the O-symbol is uniform in μ for μ in a compact interval in $\mathbf{R}_+$ containing $2/\sqrt{27}$. A general expression for the $a_n^\pm$ is available to us via the binomial theorem, but it is too unwieldy to be of intrinsic value. We note that $a_n^+ = a_n^-$ for even n.

6. Uniform expansion of P at the caustic. We have obtained expansions of P_1 and P_2. Using (2.3), (3.8), and (5.1), and restoring the large parameter x and the function f, we get the uniform approximation as $x\to\infty$:

$$
\begin{aligned}
P\left(-x,\left(\frac{2}{\sqrt{27}}-\alpha\right)x^{3/2}\right) &= e^{ix^2[f(t_2;\mu)+f(t_3;\mu)]/2}\\
&\quad\cdot\left\{p_0(\alpha)\frac{2\pi}{x^{1/6}}\cdot Ai(-x^{4/3}\zeta)\left[1+O\left(\frac{1}{x^2}\right)\right]\right.\\
&\quad\left.+\,q_0(\alpha)\frac{2\pi}{ix^{5/6}}\cdot Ai'(-x^{4/3}\zeta)\left[1+O\left(\frac{1}{x^2}\right)\right]\right\}\\
&\quad+e^{ix^2 f\left(t_1;\frac{2}{\sqrt{27}}-\alpha\right)}\sqrt{\frac{\pi}{3t_1^2-1}}\frac{1+i}{x^{1/2}}\cdot\left[1+O\left(\frac{1}{x^2}\right)\right].
\end{aligned} \tag{6.1}
$$

Here, $t_1(\mu)=t_1((2/\sqrt{27})-\alpha)=-(2/\sqrt{3})\sin(\frac{\pi}{3}+\phi)$ where $3\phi=\arcsin(1-(\sqrt{27}/2)\alpha)$. The nature of the approximation when ζ is bounded away from 0 is readily available when the asymptotic forms of the Airy function and its derivative are used [1]; recall that

$$
Ai(z)=\frac{e^{-2z^{3/2}/3}}{2\sqrt{\pi}z^{1/4}}\left(1+O\left(\frac{1}{z^{3/2}}\right)\right),\quad|\arg z|<\pi, \tag{6.2}
$$

$$
\begin{aligned}
Ai(-z) &= \frac{1}{\sqrt{\pi}z^{1/4}}\left[\sin\left(\frac{\pi}{4}+\frac{2}{3}z^{3/2}\right)\left(1+O\left(\frac{1}{z^3}\right)\right)\right.\\
&\quad\left.-\cos\left(\frac{\pi}{4}+\frac{2}{3}z^{3/2}\right)\left(\frac{5}{48z^{3/2}}+O\left(\frac{1}{z^{9/2}}\right)\right)\right],\\
&\quad|\arg z|<\frac{2\pi}{3}
\end{aligned} \tag{6.3}
$$

$$
Ai'(z)=\frac{-z^{1/4}e^{-2z^{3/2}/3}}{2\sqrt{\pi}}\left(1+O\left(\frac{1}{z^{3/2}}\right)\right),\quad|\arg z|<\pi, \tag{6.4}
$$

$$
Ai'(-z)=-\frac{z^{1/4}}{\sqrt{\pi}}\left[\cos\left(\frac{\pi}{4}+\frac{2}{3}z^{3/2}\right)\left(1+O\left(\frac{1}{z^3}\right)\right)\right.
$$

$$+ \sin\left(\frac{\pi}{4}+\frac{2}{3}z^{3/2}\right)\left(\frac{-7}{48z^{3/2}}+O\left(\frac{1}{z^{9/2}}\right)\right)\Bigg], \tag{6.5}$$

$$|\arg z| < \frac{2\pi}{3},$$

as $z \to \infty$ in the indicated sectors.

If α is positive and bounded away from 0, then so is ζ. Hence

$$\begin{aligned} x^{-1/6}Ai(-x^{4/3}\zeta) &= \frac{1}{\sqrt{\pi}\zeta^{1/4}x^{1/2}}\left\{\sin\left(\frac{\pi}{4}+\frac{2}{3}x^2\zeta^{3/2}\right)\left(1+O\left(\frac{1}{x^4}\right)\right)\right. \\ &\quad \left. - \cos\left(\frac{\pi}{4}+\frac{2}{3}x^2\zeta^{3/2}\right)\left(\frac{5}{48x^2\zeta^{3/2}}+O\left(\frac{1}{x^6}\right)\right)\right\} \end{aligned}$$

and

$$\begin{aligned} x^{-5/6}Ai'(-x^{4/3}\zeta) &= \frac{-\zeta^{1/4}}{\sqrt{\pi}x^{1/2}}\left\{\cos\left(\frac{\pi}{4}+\frac{2}{3}x^2\zeta^{3/2}\right)\left(1+O\left(\frac{1}{x^4}\right)\right)\right. \\ &\quad \left. + \sin\left(\frac{\pi}{4}+\frac{2}{3}x^2\zeta^{3/2}\right)\left(\frac{-7}{48x^2\zeta^{3/2}}+O\left(\frac{1}{x^6}\right)\right)\right\} \end{aligned}$$

by (6.3) and (6.5). Thus, (6.1) reduces to a sum of three oscillatory terms, each of order $x^{-1/2}$ as $x \to \infty$. This is what we would expect from stationary phase inside the caustic (compare (1.3)).

On the other hand, if α is negative and bounded away from 0, then $\zeta^{3/2}$ lies in a segment of the imaginary axis bounded away from the origin. Accordingly, $\arg \zeta^{3/2} = \pm\pi/2$ so that the asymptotic forms (6.2) and (6.4) show that (6.1) reduces to an oscillatory term of order $x^{-1/2}$, plus two exponentially decaying terms. This is consistent with stationary phase outside the caustic (compare (1.5)).

Finally, we examine (6.1) when $\alpha = 0$. Again from [1], we see that $Ai(0) = \Gamma(1/3)/(2\cdot 3^{1/6}\pi)$ and $Ai'(0) = -3^{1/6}\Gamma(2/3)/(2\pi)$. Since $1+i = \sqrt{2}e^{\pi i/4}$, $t_1(2/\sqrt{27}) = -2/\sqrt{3}$, and $\eta = 0$, (6.1) reduces to

$$\begin{aligned} P\left(-x, \frac{2}{\sqrt{27}}x^{3/2}\right) &= e^{-2ix^2/3+\pi i/4}\sqrt{\frac{2\pi}{3x}} + e^{ix^2/12}\frac{\Gamma(1/3)}{3^{1/3}x^{1/6}} \\ &\quad - e^{ix^2/12}\frac{i\Gamma(2/3)}{2\cdot 3^{2/3}x^{5/6}} + O\left(\frac{1}{x^{2+(1/6)}}\right) \end{aligned}$$

as $x \to +\infty$. Thus we recover (1.4). Here, use has been made of (4.5) and (4.6).

A full expansion of $P(-x, \mu x^{3/2})$ follows from the expansions of P_1 and P_2, although the coefficients become more complicated as more terms are included. The extension to complex values of x and μ (for the analytic continuation of P) can be accomplished through the expansions for P_1 and P_2 for λ, μ complex, although as applications center on real values, we have not been overly concerned with developing the full range of complex values for which our expansion is valid.

We now turn to the work in [12]. Here, the authors examine the asymptotic behaviour of the function

$$\tilde{P}(X,Y) = \int_{-\infty}^{+\infty} \exp(i(t^4 + Xt^2 + Yt))dt \tag{6.6}$$

for large values of the parameters X, Y. P and $\tilde{P}$ are related by

$$\tilde{P}(X,Y) = \frac{1}{\sqrt{2}}P\left(X, \frac{Y}{\sqrt{2}}\right).$$

In developing the expansions of $\tilde{P}(X,Y)$ away from the caustic $27Y^2+8X^3=0$, Stamnes and Spjelkavik first apply the method of stationary phase to integrals of the form

$$\int_{-\infty}^{+\infty} g(t)e^{ikh(t)}dt, \quad k\to\infty, \tag{6.7}$$

with $g(t)=1$, and $h(t)=t^4+Xt^2+Yt$. They then set $k=1$ to yield the desired large X or large Y behaviour of $\tilde{P}$; see [12, p. 1338, §3.1].

For X and Y near the caustic, the authors formally invoke the method of Chester et al. [3], and indeed, include a brief outline of the uniform asymptotic theory for integrals of the form (6.7) in an appendix. Let J_U denote the contribution to the integral (6.7) due to the coalescing stationary points (thus, J_U plays the same role as our P_1 (cf. (2.4))). The authors claim that

$$J_U \sim 2\pi e^{ik[h(t_2)+h(t_3)]}\sum_{m=0}^{2}(p_mF_m+q_mG_m) \tag{6.8}$$

as $|X|,Y\to\infty$, uniformly valid near the caustic; see [12, eq. (3.31)]. Here, $h(t)$ is the phase function of (6.7) with Y replaced by $-Y$, and t_2 and t_3 are the critical points of $h(t)$ that coalesce as (X,Y) approaches the caustic, (X,Y) remaining bounded away from the origin in the XY-plane. The functions F_m and G_m are given by

$$\begin{aligned} F_0&=k^{-1/3}Ai(-\zeta k^{2/3}) & G_0&=-ik^{-2/3}Ai'(-\zeta k^{2/3})\\ F_1&=0 & G_1&=ik^{-4/3}Ai(-\zeta k^{2/3})\\ F_2&=2k^{-5/3}Ai'(-\zeta k^{2/3}) & G_2&=2ik^{-4/3}\zeta Ai(-\zeta k^{2/3}) \end{aligned}$$

where

$$\frac{4}{3}\zeta^{3/2}=h(t_3)-h(t_2).$$

The p_m and q_m are determined similarly as in (3.7) of this paper, and are presented in the appendix of [12, eqs. (A18)–(A23)].

However, care must be taken in using expansion (6.8) to obtain the large negative-X behaviour of $\tilde{P}$ near the caustic. (There is a typographical error in the phase function $h(t_2)+h(t_3)$, which should likely be half the stated value.) Throughout the appendix and §2 of [12], k appears as a large positive parameter. Yet, in several places, k is set equal to one prior to examining the large X behaviour of (6.7); see, for instance, equations (3.5), (5.3) and the discussion in §3.2 and note the absence of k in expansion (3.30). This naturally leads one to suspect that the same is being done in (6.8), although no explicit mention of this is made in [12]. It should be pointed out that (6.8) is that part of the expansion of

$$P^*(X,Y;k)=\int_{-\infty}^{+\infty}\exp(ik(t^4+Xt^2+Yt))dt$$

due to the coalescing saddles of the phase function t^4+Xt^2+Yt, and not part of an expansion of $\tilde{P}$. However, expansion (6.8) can be used to deduce the large negative-X behaviour of $\tilde{P}$ near the caustic via the relation

$$\tilde{P}(X,Y)=k^{1/4}P^*(\bar{X},\bar{Y};k),$$

where k is a large parameter, $X = k^{1/2}\bar{X}$ and $Y = k^{3/4}\bar{Y}$. Note that $27Y^2 + 8X^3 = k^{3/2}(27\bar{Y}^2 + 8\bar{X}^3)$; thus, if (X, Y) is near the caustic, so is $(\bar{X}, \bar{Y})$.

The integral P^* is also related to the Pearcey function $P(x, y)$ given in (1.1) by

$$P^*(X, Y; k) = 2^{-1/2}k^{-1/4}P(k^{1/2}X, 2^{-1/2}k^{3/4}Y).$$

By setting $x = -k^{1/2}X$ and $\mu = 2^{-1/2}(-X)^{-3/2}Y$ in our expansion for $P_1(x^2; \mu)$ (cf. (3.8)), we obtain the first three terms in (6.8). Thus, modulo misprints, (6.8) appears correct.

Finally, we turn to the "transitional approximation" developed in [12]. This was used in the numerical evaluation of $\tilde{P}(X, Y)$ for (X, Y) in a band covering the caustic extending 0.05 units in the X-direction on either side of the caustic, with $Y > 0$ (see [12, p. 1349, second paragraph]). In this calculation, it was assumed that $X^2 + Y^2 > 16$ and $|X|, Y \leq 8$. The "transitional approximation" developed is an asymptotic approximation of J_T in the immediate vicinity of the caustic [*sic*], where J_T represents the contribution to (6.6) due to coalescing stationary points of $h(t)$. The authors assert that

$$(6.9)\quad J_T \sim 2\pi d \cdot \exp(ih_0)\left[Ai(z) + icAi^{(4)}(z) - \frac{c^2}{2}Ai^{(8)}(z) - \frac{ic^3}{6}Ai^{(12)}(z) + \cdots\right]$$

where

$$\begin{aligned} h_0 &= -\tfrac{5}{36}X^2 + Y(\tfrac{-X}{6})^{1/2} & h_1 &= -8[\tfrac{Y}{8} - (\tfrac{-X}{6})^{3/2}] \\ h_3 &= -4(\tfrac{-X}{6})^{1/2} & h_4 &= 1 \\ d &= (3|h_3|)^{-1/3} & \epsilon &= \operatorname{sgn}(h_3) \\ z &= \epsilon h_1 d & c &= h_4 d^4 \end{aligned}$$

see [12, p. 1343, eq. (3.30)].

There are two points which we wish to make regarding the preceding approximation. First, the derivation is purely formal with no mention being made of the region of validity. Second, the authors did not actually use the expansion in the form given in (6.9), but instead used one in which each of the Airy functions $Ai^{(4j)}$, $j = 0, 1, 2, 3$, is replaced by the first term in its Maclaurin expansion; cf. the first three lines on p. 1344 of [12]. That is, the authors replace (6.9) by

$$(6.10)\qquad J_T = 2\pi d \cdot e^{ih_0}\left\{\sum_{j=0}^{3} c^j Q_j(z) + O(c^4)\right\},$$

where each Q_j is an expression of the form α or βz, α and β being constants.

In order for this to be valid, the implied O-terms in (6.10) involving z, resulting from approximating the $Ai^{(4j)}$ by the first terms of their Maclaurin series, must be $O(c^4)$ for large $|X|$. In particular, we must have $z = O(c^4)$. Since $z = -h_1 c^{1/4}$, this is equivalent to

$$h_1 = O(c^{15/4})$$

or, upon restoring X and Y,

$$Y - 8(-X/6)^{3/2} = O((-X)^{-5/2})$$

whence

$$27Y^2 + 8X^3 = O((-X)^{-1}).$$

This displays a condition on how quickly the point (X, Y) must approach the caustic.

In closing, we note that our work in developing a uniform expansion of $P(x, y)$ near the caustic is easily extended to that of deriving uniform expansions of $\partial P/\partial x$ and $\partial P/\partial y$ near the caustic for a range of complex x and y.

Acknowledgments. The author is indebted to Professor R. Wong for his guidance throughout the development of this work, and to Professor F. Ursell for detecting an error in an earlier version of this paper and for providing comments that greatly improved the discussion in §6.

REFERENCES

[1] M. ABRAMOVITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover Publications, New York, 1972.

[2] N. BLEISTEIN AND R. A. HANDELSMAN, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, New York, 1975.

[3] C. CHESTER, B. FRIEDMAN,AND F. URSELL, *An extension of the method of steepest descents*, Proc. Cambridge Phil. Soc., 53 (1957), pp. 599–611.

[4] E. T. COPSON, *Asymptotic Expansions*, Cambridge Tracts in Math. and Math. Phys. 55, Cambridge Univ. Press, London, New York, 1965.

[5] J. N. L. CONNOR AND D. FARRELLY, *Molecular collisions and cusp catastrophes: Three methods for the calculation of Pearcey's integral and its derivatives*, Chem. Phys. Letters, 81, (1981), pp. 306–310.

[6] R. GILMORE, *Catastrophe Theory for Scientists and Engineers*, John Wiley, New York, 1981.

[7] V. GUILLEMIN AND S. STERNBERG, *Geometric Asymptotics*, Math. Surveys, Amer. Math. Soc., Providence, RI, 1977.

[8] R. HABERMAN AND R.-J. SUN, *Nonlinear cusped caustics for dispersive waves*, Studies in Applied Math., to appear.

[9] V. P. MASLOV AND M. V. FEDORIUK, *Semi-Classical Approximation in Quantum Mechanics*, Kluwer, Boston, 1981.

[10] R. MIURA, *Explicit roots of the cubic polynomial and applications*, CMS Applied Math. Notes, 5 (1980), pp. 22–40.

[11] T. PEARCEY, *The structure of an electromagnetic field in the neighborhood of a cusp of a caustic*, Lond. Edinb. Dubl. Phil. Mag., 37 (1946), pp. 311–317.

[12] J. J. STAMNES AND B. SPJELKAVIK, *Evaluation of the field near a cusp of a caustic*, Optica Acta, 30 (1983), pp. 1331–1358.

[13] C. UPSTILL, *et al., The double-cusp unfolding of the* ${}^0X_9$ *diffraction catastrophe*, Optica Acta, 29 (1982), pp. 1651–1676.

[14] F. URSELL, *Integrals with a large parameter. Several nearly coincident saddle points*, Proc. Camb. Phil. Soc., 72 (1972), pp. 49–65.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 1006–1018, July 1989

013

INFINITE INTEGRALS INVOLVING THREE SPHERICAL BESSEL FUNCTIONS*

A. GERVOIS†‡ AND H. NAVELET†

Abstract. The integrals $\int_0^\infty t^n j_{l_1}(at) j_{l_2}(bt) j_{l_3}(ct)\, dt$ are calculated, where j_l is the spherical Bessel function for any integer indices n, l_1, l_2, l_3 and real positive parameters a, b, c using two different methods. In the first, $n+l_1+l_2+l_3$ must be an even integer but the techniques may be generalized to ordinary Bessel functions with *noninteger* indices. The second method does not depend on the parity of $n+l_1+l_2+l_3$ but remains valid only for integer indices.

Key words. spherical Bessel functions, infinite integrals, Appell functions

AMS(MOS) subject classifications. 33A40, 44A15, 33A35

1. Introduction.

Nuclear physicists [1] are often faced with integrals of the form

$$\int_0^\infty t^n j_{l_1}(at) j_{l_2}(bt) j_{l_3}(ct)\, dt \tag{1.1}$$

where l_1, l_2, l_3 are positive (or zero) integers and the integer n satisfies convergence conditions

$$-(l_1+l_2+l_3) \leqq n \leqq 2.$$

Parameters a, b, c are real positive and the $j_l(x)$ are the spherical Bessel functions

$$j_l(x) = \sqrt{\pi/2x} J_{l+1/2}(x). \tag{1.2}$$

A calculation of (1.1) has already been performed, in the frame of coupled channel theory [2], but with the following restrictions:

(1.3)
- (a) $n = 2 - M$, M is an even integer such that $0 \leqq M \leqq l_1+l_2+l_3-2l_M$, $l_M = m_{ax}(l_1, l_2, l_3)$.
- (b) $l_1+l_2+l_3$ is even.
- (c) $|l_1-l_2| < l_3 < l_1+l_2$.
- (d) The parameters a, b, c obey the triangular inequality, namely

$$|a-b| < c < a+b.$$

Restrictions (b) and (c) correspond to conservation of parity and angular momentum, respectively. They allow the use of spherical harmonics and Clebsch-Gordon coefficients. As to $n=2$, this condition simply says that $t^2\, dt$ is the volume element in the three-dimensional space. The result is written as a sum of expressions

$$\begin{pmatrix} l_1 & l_2 & l_3 \\ m & -m & 0 \end{pmatrix} P_{l_1}^m(\cos\theta_{13}) P_{l_2}^{-m}(\cos\theta_{23})$$

where the first factor is a Clebsch-Gordon and the $P_l^{\pm m}$ are Legendre polynomials. Angles θ_{13}, θ_{23} may be defined only because of the peculiar geometry of the system (condition (d)). In spite of its elegance, the proof of [2] remains specific to half-integer Bessel functions and needs crucially all conditions listed above.

* Received by the editors January 21, 1987; accepted for publication (in revised form) October 4, 1988.

† Service de Physique Théorique, CEN-Saclay, 91191 Gif-sur-Yvette Cedex, France.

‡ Groupe de Physique Cristalline, Unité 804 Associée au CNRS, Université de Rennes I, Campus de Beaulieu, 35042 Rennes Cedex, France.

Actually, there exists a formal expression [3] for the integral

$$I(\lambda, \mu, \nu, \rho) \equiv \int_0^\infty t^{\lambda-1} J_\mu(at) J_\nu(bt) H_\rho^{(1)}(ct)\, dt \tag{1.4}$$

provided the convergence conditions

$$\mathrm{Re}\,(\lambda) < \tfrac{5}{2}, \quad \mathrm{Re}\,(\lambda+\mu+\nu-|\rho|) > 0 \quad (\mathrm{Re}\,(\lambda+\mu+\nu+\rho) > 0 \text{ for the integral where } H_\rho^{(1)}(ct) \text{ is replaced by } J_\rho(ct)), \tag{1.5}$$

hold for the indices λ, μ, ν, ρ.

We have

$$I(\lambda, \mu, \nu, \rho) = \frac{1}{2\pi} e^{i\pi\beta} \left(\frac{2}{c}\right)^\lambda \left(\frac{a}{c}\right)^\mu \left(\frac{b}{c}\right)^\nu \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\gamma)\Gamma(\gamma')} \times F_4\left(\alpha, \beta, \gamma, \gamma'; \frac{a^2}{c^2}, \frac{b^2}{c^2}\right) \tag{1.6a}$$

with

$$\alpha = \frac{\lambda+\mu+\nu+\rho}{2}, \quad \beta = \frac{\lambda+\mu+\nu-\rho}{2}, \quad \gamma = \mu+1, \quad \gamma' = \nu+1 \tag{1.6b}$$

and F_4 is the Appell [4] function, defined as a double series in the convergence domain $|c| > |a| + |b|$.

However, this result is not useful because the function F_4 cannot be calculated easily and the analytic continuation is not possible when a, b, c may be considered as the sides of a triangle.

Simplifications occur when F_4 factorizes into functions of one variable that allows an analytic continuation in a, b, c. These cases have been listed long ago [5a], [5b], and in previous works [6a], [6b], [7a], [7b], we have used this factorization to calculate numerous integrals of type (1.4) when λ, μ, ν, ρ are related by one or two linear relations. By using recurrence relations between Bessel functions it is possible to enlarge the number of integrals of type (1.4) that can be computed analytically. These new integrals correspond to λ, μ, ν, ρ values differing from the previous ones by an integer. This integer is always positive for the parameter λ. As we will see in § 2, this general method is a powerful tool to get as a byproduct integrals of type (1.1) but with the restriction $l_1+l_2+l_3+n$ even. In § 3, we use a direct method, specific of the spherical Bessel functions [1], but which removes the restriction $l_1+l_2+l_3+n$ even.

2. First method. We consider here integrals $I(\lambda, \mu, \nu, \rho)$, (1.4) and in a first step do not impose any restriction on the indices except the convergence conditions (1.5). In § 2.1, we show how recurrence relations between Bessel functions lead to recurrence relations between contiguous integrals $I(\lambda, \mu, \nu, \rho)$. We then derive a whole class of integrals that can be calculated from already known integrals in an explicit form (§ 2.2). In the last section, § 2.3, we apply these results to integrals (1.1).

2.1. Recurrence relations. The well-known recurrence relation between three contiguous (nonmodified) Bessel functions

$$t[Z_{\sigma+1}(t) + Z_{\sigma-1}(t)] = 2\sigma Z_\sigma(t) \tag{2.1}$$

where $Z_\sigma = J_\sigma$, Y_σ or $H_\sigma^{(1)}$ leads to a straightforward relation between contiguous integrals. Setting $t = cx$, $\sigma = \rho - 1$, we have for any λ, μ, ν, ρ,

$$I(\lambda+2, \mu, \nu, \rho) + I(\lambda+2, \mu, \nu, \rho-2) = \frac{2}{c}(\rho-1) I(\lambda+1, \mu, \nu, \rho-1). \tag{2.2}$$

Similarly, the differential relations

$$\left(\frac{\sigma}{z} \pm \frac{\partial}{\partial z}\right) Z_\sigma(tz) = t Z_{\sigma\mp1}(tz) \tag{2.3}$$

together with integration by parts yield other (less obvious) relations. For example, we will need

$$I(\lambda+1,\mu,\nu+1,\rho)=\frac{b}{c}I(\lambda+1,\mu,\nu,\rho-1)-\frac{a}{c}I(\lambda+1,\mu+1,\nu+1,\rho-1)$$
$$-\frac{2+\nu-\mu-\lambda-\rho}{c}I(\lambda,\mu,\nu+1,\rho-1). \tag{2.4}$$

We now consider the case

$$\nu-\mu=q,\qquad \lambda+\mu+\rho-\nu=2(k+1),$$
$$\mathcal{J}_{\mu\ \rho}^{q\ k}\equiv\int_0^\infty dt\,t^{1-\rho+q+2k}J_\mu(at)J_{\mu+q}(bt)H_\rho^{(1)}(ct) \tag{2.5}$$

where q, k are positive or zero integers. The justification of this restriction will appear clearly later. For these integrals, (2.2) reads

$$\mathcal{J}_{\mu\ \rho}^{q\ k+1}=-\mathcal{J}_{\mu\ \rho-2}^{q\ k}+\frac{2}{c}(\rho-1)\mathcal{J}_{\mu\ \rho-1}^{q\ k}$$

whence, by recursion

$$\mathcal{J}_{\mu\ \rho}^{q\ k}=\sum_{p=0}^{k}C_k^p\left(\frac{2}{c}\right)^{k-p}(-)^p\frac{\Gamma(\rho-p)}{\Gamma(\rho-k)}\mathcal{J}_{\mu\ \rho-k-p}^{q\ 0}$$

and (2.4) becomes

$$\mathcal{J}_{\mu\ \rho}^{q+1\ 0}=\frac{b}{c}\mathcal{J}_{\mu\ \rho-1}^{q\ 0}-\frac{a}{c}\mathcal{J}_{\mu+1\ \rho-1}^{q\ 0}$$

because the last term in the right-hand side cancels out. By recursion again, we get

$$\mathcal{J}_{\mu\ \rho}^{q\ 0}=\sum_{m=0}^{q}C_q^m(-)^m\left(\frac{b}{c}\right)^{q-m}\left(\frac{a}{c}\right)^m\mathcal{J}_{\mu+m\quad\rho-q}^{0\ 0}.$$

This allows us to write $\mathcal{J}_{\mu\ \rho}^{q\ k}$ as a finite linear combination of known integrals $\mathcal{J}_{\mu'\ \rho'}^{0\ 0}$ [5]

$$\mathcal{J}_{\mu\ \rho}^{q\ k}=\sum_{p=0}^{k}C_k^p\left(\frac{2}{c}\right)^{k-p}(-)^p\frac{\Gamma(\rho-p)}{\Gamma(\rho-k)}\sum_{m=0}^{q}C_q^m(-)^m\left(\frac{b}{c}\right)^{q-m}\left(\frac{a}{c}\right)^m\mathcal{J}_{\mu'\ \rho'}^{0\ 0} \tag{2.6a}$$

with

$$\mu'=\mu+m\quad(m=0,1,\cdots,q),\qquad \rho'=\rho-k-p-q\quad(p=0,1,\cdots,k) \tag{2.6b}$$

and

$$\mathcal{J}_{\mu'\ \rho'}^{0\ 0}\equiv\int_0^\infty t^{1-\rho'}J_{\mu'}(at)J_{\mu'}(bt)H_{\rho'}^{(1)}(ct)\,dt. \tag{2.7}$$

Note that $\rho'=(k-p)+(\rho-2k-q)$ is positive for $k>p$. Indeed, the convergence condition at $t=\infty$ implies $\rho-q-2k>-\frac{1}{2}$.

2.2. New integrals. The general expression (1.6) for $I(\lambda,\mu,\nu,\rho)$ is of particular interest when F_4 factorizes into functions of one variable (in general, ${}_2F_1$ functions). These factorizations [5a], [5b] roughly fall into three cases:

(i) $\lambda=1$, any μ, ν, ρ.

(ii) $\lambda=\nu+2$, $\mu=\pm\rho$ (and $\lambda=\mu+2$, $\nu=\pm\rho$).

(iii) $\lambda=2\pm\rho$, $\mu=\nu$.

Some of the corresponding integrals may be found in the usual table handbooks [8], [9]. We calculate others in recent papers [6a], [6b], [7a], [7b]. In case (i) F_4 is the product of two ${}_2F_1$ functions in intermediate variables and recurrence relations (2.4)-(2.6) then simply say that when λ is a (strictly) positive integer the integral $I(\lambda, \mu, \nu, \rho)$—or the corresponding F_4 function—is now a finite sum of such products of ${}_2F_1$ functions in the same variable. However, up to now, there is no simple close formula for such integrals in the more general case; we have seen some peculiar cases in previous publications [10].

Cases (ii) and (iii) are partially the same. In the following, we will drop the configuration of indices $\lambda = 2+\nu$, $\mu = -\rho$ although the same technique probably works in that case too. For the remaining possibilities, without any restriction, we may assume $\lambda = 2-\rho$, $\mu = \nu$ and we get then precisely the integral $\mathcal{J}^{0\ 0}_{\mu\ \rho}$ of definition (2.7). Here we recall the corresponding factorization for F_4:

$$F_4\left(\mu+1, \mu+1-\rho, \mu+1, \mu+1; -\frac{X}{(1-X)(1-Y)}, -\frac{Y}{(1-X)(1-Y)}\right)$$
$$= [(1-X)(1-Y)]^{\mu+1-\rho}\, {}_2F_1(\mu+1-\rho, 1-\rho, \mu+1; XY),$$

which yields for $\mathcal{J}^{0\ 0}_{\mu\ \rho}$ the final expression [6a]:

$$\text{(2.8a)} \qquad \mathcal{J}^{0\ 0}_{\mu'\ \rho'} = \frac{i}{\pi}\sqrt{\frac{2}{\pi}}\frac{(ab)^{\rho'-1}}{c^{\rho'}}(sh\tilde{U}_c)^{\rho'-1/2}\, e^{i\pi(\rho'-1/2)}\mathcal{Q}^{-\rho'+1/2}_{\mu'-1/2}(ch\tilde{U}_c),$$

$$c < |a-b|,$$

$$\text{(2.8b)} \qquad = \frac{i}{\pi}\sqrt{\frac{2}{\pi}}\frac{(ab)^{\rho'-1}}{c^{\rho'}}(shU_c)^{\rho'-1/2}\, e^{i\pi(\mu'-1/2)}\mathcal{Q}^{-\rho'+1/2}_{\mu'-1/2}(chU_c),$$

$$c > a+b,$$

$$\text{(2.8c)} \qquad = \frac{1}{\pi}\sqrt{\frac{2}{\pi}}\frac{(ab)^{\rho'-1}}{c^{\rho'}}(\sin\varphi_c)^{\rho'-1/2}\left\{\frac{\pi}{2}P^{-\rho'+1/2}_{\mu'-1/2}(\cos\varphi_c) - iQ^{-\rho'+1/2}_{\mu'-1/2}(\cos\varphi_c)\right\},$$

$$|a-b| < c < a+b.$$

The hyperbolic angles $\tilde{U}_c$, U_c relative to c when a, b, c do not form a triangle are defined by

$$\text{(2.9a)} \qquad c^2 = a^2+b^2-2ab\, ch\tilde{U}_c \quad \text{if } c < |a-b|,$$

$$\text{(2.9b)} \qquad c^2 = a^2+b^2+2ab\, chU_c \quad \text{if } c > a+b.$$

In the triangle configuration, φ_c is the true angle:

$$\text{(2.9c)} \qquad c^2 = a^2+b^2-2ab\cos\varphi_c, \qquad |a-b| < c < a+b$$

and similar relations hold for the angles or hyperbolic angles relative to a and b. The $\mathcal{P}^\tau_\sigma$, $\mathcal{Q}^\tau_\sigma$ (respectively, P^τ_σ, Q^τ_σ) are the Legendre functions outside the cut ((2.8a) and (2.8b)) (respectively, on the cut ((2.8c)) (see, for example, [11]). We must point out that $\mathcal{Q}^\tau_\sigma$ is not defined when $\sigma+\tau$ is a negative integer; however, a definition by continuation is possible when $\tau = -r$ is an integer too, using relation $\mathcal{Q}^{-r}_\sigma = (-)^r(\Gamma(\sigma-r+1)/\Gamma(\sigma+r+1))\mathcal{Q}^r_\sigma$. This will be useful later. Note that $e^{i\pi\tau}\mathcal{Q}^{-\tau}_\sigma$ is real when z is real and greater than one, so that (2.8a) is pure imaginary and $\int t^{1-\rho'}J_{\mu'}J_{\mu'}J_{\rho'}$ is zero when $c < |a-b|$ and this holds by analytic continuation when $\mu'-\rho'$ is a negative integer. If $c > a+b$ and $\mu'-\rho'$ is a nonnegative integer, $\int t^{1-\rho'}J_{\mu'}J_{\mu'}J_{\rho'}$ is also zero because of the $\sin(\mu'-\rho')\pi$ factor. When $\mu'-\rho'$ is a negative integer, the limit exists but is not zero. This will be seen in more detail in the next section.

Now, any integral $\mathcal{J}_{\mu\ \rho}^{q\ k}$ where the indices differ by integer number is rewritten as a finite sum of such Legendre functions using summation (2.6). We give here some examples (Re is the real part).

(i) Sine transforms (or cosine transforms)

$$\int_0^\infty t^{q+2k} J_\mu(at) J_{\mu+q}(b) \operatorname{sinct} dt = \sqrt{\pi c/2}\ \mathrm{Re}\,(\mathcal{J}_{\mu\ 1/2}^{q\ k}).$$

(ii) For $\rho = \nu - \mu$ and $k = 0$, we recover results calculated by another method in [10] for $\lambda = 2$ and any real μ, ν.

(iii) Bessel J_0 (or $J_1 \cdots$) transforms

$$\int_0^\infty t^{1+q+2n} J_\nu(at) J_{\nu+q}(bt) J_0(ct)\, dt = \mathrm{Re}\,(\mathcal{J}_{\nu\ 0}^{q\ n}).$$

(iv) Integrals with fractional indices such as that involving the Airy function $J_{1/3}$, for instance,

$$\int_0^\infty t^{2n+1} J_{1/3}(at) J_{1/3}(bt) J_0(ct)\, dt = \mathrm{Re}\,(\mathcal{J}_{1/3\ 0}^{0\ n})$$

or

$$\int_0^\infty t^{2/3+q} J_{1/3}(ct) J_\mu(at) J_{\mu+q}(bt)\, dt = \mathrm{Re}\,(\mathcal{J}_{\mu\ 1/3}^{q\ 0}).$$

2.3. Application to spherical Bessel functions. Let us come back to integrals (1.1). We assume without any loss of generality that $l_3 \geqq l_2 \geqq l_1$ and we set $l_2 = l_1 + q$. The correspondence with integrals $I(\lambda, \mu, \nu, \rho)$ or $\mathcal{J}_{\mu\ \rho}^{q\ k}$ obviously gives $\mu = l_1 + \frac{1}{2}$, $\nu = l_2 + \frac{1}{2}$, $\rho = l_3 + \frac{1}{2}$, $\lambda = 2 - \rho + q + 2k = n - \frac{1}{2}$, and

$$\begin{aligned} &\int_0^\infty t^n j_{l_1}(at) j_{l_2}(bt) j_{l_3}(ct)\, dt \\ &= \left(\frac{\pi}{2}\right)^{3/2} \frac{1}{\sqrt{abc}} \int_0^\infty t^{n-3/2} J_{l_1+1/2}(at) J_{l_2+1/2}(bt) J_{l_3+1/2}(ct)\, dt \\ &= \left(\frac{\pi}{2}\right)^{3/2} \frac{1}{\sqrt{abc}}\, \mathrm{Re}\left\{ \mathcal{J}_{l_1+1/2, l_3+1/2}^{l_2-l_1, (n-2+l_3-l_2+l_1)/2} \right\} \end{aligned} \tag{2.10}$$

where Re denotes the real part, provided

$$n + l_1 + l_2 + l_3 \quad \text{is even}, \tag{2.11a}$$

$$n - 2 + l_3 - l_2 + l_1 \geqq 0. \tag{2.11b}$$

The first condition is actually fulfilled in nuclear physics problems and is related to the conservation of parity. Condition (2.11b) is automatically satisfied for $n = 2$ or $n = 1$ as $l_3 \geqq l_2 \geqq l_1$ and $l_1 + l_2 + l_3$ has a given parity. For $n = 0$, cases $l_1 = 0$, $l_2 = l_3 = l$ cannot be treated directly by this method. A derivation in the spirit of this paper is given in Appendix A. Note then, that the integral is a pure Fourier transform, and the result when a, b, c do not form a triangle may be found in handbooks of tables (see, for instance, [8]).

Now, when conditions (2.11) are fulfilled, using (2.6)-(2.8) with $2k = n - 2 + l_3 - l_2 + l_1$, $q = l_2 - l_1$, we get the result as a sum of integrals (2.8) with μ', ρ' half integers ($\mu' = l_1 + m + \frac{1}{2} = l_1' + \frac{1}{2}$, $\rho' = l_3 + \frac{1}{2} - k - p - q = l_3' + \frac{1}{2}$). Note that $l_3' \geqq 0$,

although it is not clear in this form. For the sake of completeness, we rewrite (2.10) when using (2.6):

$$\int_0^\infty t^n j_{l_1}(at) j_{l_2}(bt) j_{l_3}(bt)\, dt$$

$$= \left(\frac{\pi}{2}\right)^{3/2} \frac{1}{\sqrt{abc}} \sum_{p=0}^{k} C_k^p \left(\frac{2}{c}\right)^{k-p} (-)^p \frac{\Gamma(l_3+\frac{1}{2}-p)}{\Gamma(l_3+\frac{1}{2}-k)} \tag{2.12a}$$

$$\cdot \sum_{m=0}^{q} C_q^m (-)^m \left(\frac{b}{c}\right)^{q-m} \left(\frac{a}{c}\right)^m \operatorname{Re} \mathcal{J}^0_{l_1'+1/2,\, l_3'+1/2},$$

$$l_3 \geqq l_2 \geqq l_1, \quad q = l_2 - l_1, \quad 2k = n - 2 + l_3 - l_2 + l_1, \tag{2.12b}$$

$$l_1' = l_1 + m, \qquad l_3' = l_3 - k - p - q, \tag{2.12c}$$

and

$$\operatorname{Re} \mathcal{J}^0_{l_1'+1/2,\, l_3'+1/2} \equiv \int_0^\infty t^{1/2-l_3'} J_{l_1'+1/2}(at) J_{l_1'+1/2}(bt) J_{l_3'+1/2}(ct)\, dt$$

$$= 0 \quad \text{if } c < |a-b|$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{ab} \left(\frac{ab}{c}\right)^{l_3'+1/2} (\sin\varphi_c)^{l_3'} P_{l_1'}^{-l_3'}(\cos\varphi_c)$$

$$|a-b| < c < a+b \tag{2.13}$$

$$= 0 \quad \text{for } c > a+b \text{ and } l_1' \geqq l_3'$$

$$= (-)^{l_3'} \sqrt{\frac{2}{\pi}} \frac{1}{ab} \left(\frac{ab}{c}\right)^{l_3'+1/2} \frac{(shU_c)^{l_3'}}{\Gamma(l_1'+l_3'+1)\Gamma(l_3'-l_1')} \mathcal{Q}_{l_1'}^{l_3'}(chU_c)$$

$$\text{for } c > a+b \text{ and } l_1' < l_3'.$$

As l_1', l_3' are positive integers, $(\sin\varphi_c)^{l_3'} P_{l_1'}^{-l_3'}(\cos\varphi_c)$ is a polynomial in $\cos\varphi_c$ of degree $l_1'+l_3'$ whereas $(shU_c)^{l_3'} \mathcal{Q}_{l_1'}^{l_3'}(chU_c)$ is a polynomial in chU_c of degree $l_1'+l_3'-1$ when $l_3' > l_1'$ (see Appendix B).

At this stage, some comments are in order.

(i) It is worth noting that there is a shorter way to get the integrals of type (2.10) when a, b, and c do not obey the triangular inequalities.

Indeed if $c > a+b$

$$\operatorname{Re} I\left(n-\frac{1}{2}; l_1+\frac{1}{2}, l_2+\frac{1}{2}, l_3+\frac{1}{2}\right)$$

$$= \frac{\frac{1}{2}\left(\frac{2}{c}\right)^{n-1/2} \left(\frac{a}{c}\right)^{l_1+1/2} \left(\frac{b}{c}\right)^{l_2+1/(2\times\Gamma)} \left(\frac{l_1+l_2+l_3+n+1}{2}\right)}{\Gamma\left(l_1+\frac{3}{2}\right)\Gamma\left(l_2+\frac{3}{2}\right)\Gamma\left(1-\left(\frac{l_1+l_2+n-l_3}{2}\right)\right)}$$

$$\times F_4\left(\frac{l_1+l_2+l_3+n+1}{2}, \frac{l_1+l_2+n-l_3}{2}, l_1+\frac{3}{2}, l_2+\frac{3}{2}; \frac{a^2}{c^2}, \frac{b^2}{c^2}\right)$$

and similar formulae for $a > c+b$ ($c \leftrightarrow a$ and $l_3 \leftrightarrow l_1$) or $b > a+c$ ($c \leftrightarrow b$ and $l_3 \leftrightarrow l_2$). Then for $n+l_1+l_2+l_3$ even, $(l_1+l_2+n-l_3)/2$ is an integer and the integral vanishes

if $l_1+l_2+n-l_3 \geqq 2$. In the other configuration ($l_1+l_2+n-l_3 \leqq 0$) the F_4 function is a polynomial of degree $(l_3-(l_1+l_2+n))/2=p$ in the variable a^2/c^2 and b^2/c^2 of the form

$$F_4\left(\alpha, -p, \gamma, \gamma', \frac{a^2}{c^2}, \frac{b^2}{c^2}\right) = \sum_{k=0}^{p} (\alpha)_k (-p)_k \sum_{m=0}^{k} \frac{\left(\frac{a^2}{c^2}\right)^m \left(\frac{b^2}{c^2}\right)^{k-m}}{(\gamma)_m (\gamma')_{k-m} m!(k-m)!}$$

where $(\beta)_q = \Gamma(\beta+q)/\Gamma(\beta)$ is the Pochammer symbol.

For instance, the case $l_1+l_2+n=l_3$ leads to a very simple formula since the F_4 function reduces to one (for reasons of convergence $l_3 \leqq l_1+l_2+2$):

$$\int_0^\infty t^{l_3-(l_1+l_2)} j_{l_1}(at) j_{l_2}(bt) j_{l_3}(ct)\, dt$$

$$= \frac{\pi^{3/2}}{2^{3+l_1+l_2-l_3}} \frac{a^{l_1} b^{l_2}}{c^{l_3+1}} \times \frac{\Gamma(l_3+\frac{1}{2})}{\Gamma(l_1+\frac{3}{2})\Gamma(l_2+\frac{3}{2})}, \qquad c > a+b.$$

Note that this result is valid in the general case when $(\lambda+\mu+\nu-\rho)/2$ is an integer. Thus we have the well-known formula:

$$\int_0^\infty t^{\rho-\nu-\mu-1} J_\mu(at) J_\nu(bt) J_\rho(ct)\, dt = \frac{1}{2} \frac{\left(\frac{a}{2}\right)^\mu \left(\frac{b}{2}\right)^\nu \Gamma(\rho)}{\left(\frac{c}{2}\right)^\rho \Gamma(\nu+1)\Gamma(\mu+1)}, \qquad c > a+b.$$

Unfortunately, this method does not apply when a, b, c obey the triangular inequalities because, as emphasized before, the analytic continuation of the F_4 function is unknown in the general case.

(ii) The proof above holds when $n \geqq 0$ and is still valid when $n<0$, provided $n-2+l_3-l_2+l_1 \geqq 0$.

It is easy to show now that it also holds when $n-2+l_3-l_2+l_1<0$, i.e., for any negative integer n provided the convergence condition at $t=0$ $(n+l_1+l_2+l_3) \geqq 0$ holds.

Let us show it briefly. We set $m=|n|=-n$. As $m \leqq l_1+l_2+l_3$, we may split m into three terms $m=m_1+m_2+m_3$ such that m_i is an integer and $0 \leqq m_i \leqq l_i$. Now, using m_i times the recurrence relation (2.1), we replace j_{l_i}/t^{m_i} by a linear combination of m_i+1 spherical Bessel functions $j_{l_i'}$ ($l_i'=l_i+m_i,\ l_i+m_i-2, \cdots, l_i-m_i+2,\ l_i-m_i$) and the whole integrand is now a sum of products of spherical Bessel functions with no power term. For example, when we use

$$\frac{j_1(at)}{t} = \frac{a}{3}[j_2(at)+j_0(at)], \qquad \frac{j_1(bt)}{t} = \frac{b}{3}[j_2(bt)+j_0(bt)],$$

$$\frac{j_2(ct)}{t^2} = c^2\left[\frac{j_4(ct)}{35} + \frac{2j_2(ct)}{21} + \frac{j_0(ct)}{15}\right]$$

the integral

$$\int_0^\infty \frac{j_1(at) j_1(bt) j_2(ct)}{t^4}\, dt$$

is a sum of known integrals $\int_0^\infty j_{l_1'}(at) j_{l_2'}(bt) j_{l_3'}(ct)\, dt$ with l_1', $l_2'=0, 2$; $l_3'=0, 2, 4$.

To summarize, we have enlarged the conditions (1.3) that were necessary in the proof of [2] to the case where a, b, and c, respectively (l_1, l_2, and l_3) do not obey the triangular inequalities and where n is no longer restricted to be equal to two. The result is given in terms of a linear combination of Legendre functions of one angle or

pseudo-angle φ_c, U_c, or $\hat{U}_c$ instead of products of Legendre polynomials in φ_a and φ_b in the triangular case (see Table 1). The method is general and applies also when the index of the Bessel functions is noninteger.

TABLE 1

Expression of the integral with three spherical Bessel functions in terms of the Legendre functions when $n+l_1+l_2+l_3$ *is even.*

$$I=\int_0^\infty dt\, t^n j_{l_1}(at)j_{l_2}(bt)j_{l_3}(ct), \qquad n\geqq 2, \quad n+l_1+l_2+l_3\geqq 0.$$

For $n<0$, we come back to integrals with $n=0$ using recurrence (2.1). For $n=l_1=0$ and $l_2=l_3=l$ see Appendix equations (A1)-(A2). In all other cases, if $l_3\geqq l_2\geqq l_1$

$$I=\left(\frac{\pi}{2}\right)^{3/2}\frac{1}{\sqrt{abc}}\,\mathrm{Re}\,\mathcal{J}^{q\ k}_{l_1+1/2\ l_3+1/2}$$

where $q=l_2-l_1$, $k=(n-2+l_3-l_2+l_1)/2$ are positive or zero integers,

$$\mathcal{J}^{q\ k}_{l_1+1/2\ l_3+1/2}=\sum_{p=0}^{k}C_k^p\left(\frac{2}{c}\right)^{k-p}(-)^p\frac{\Gamma(l_3+\tfrac{1}{2}-p)}{\Gamma(l_3+\tfrac{1}{2}-k)}\sum_{m=0}^{q}C_q^m(-)^m\left(\frac{b}{c}\right)^{q-m}\left(\frac{a}{c}\right)^m\mathcal{J}^{0}_{l_1'+1/2\ l_3'+1/2}$$

$$l_1'=l_1+m, \quad l_3'=l_3-k-p-q, \quad (l_3'\geqq 0)$$

and

$$\mathcal{J}^{0}_{l_1'+1/2\ l_3'+1/2}=0 \quad \text{if } c<|a-b|$$

$$=0 \quad \text{if } c>a+b \text{ and } l_1'\geqq l_3'$$

$$=(-)^{l_3'}\sqrt{\frac{2}{\pi}}\frac{1}{ab}\left(\frac{ab}{c}\right)^{l_3'+1/2}\frac{(shU_c)^{l_3'}}{\Gamma(l_1'+l_3'+1)\Gamma(l_3'-l_1')}\mathcal{Q}_{l_1'}^{l_3'}(chU_c) \quad \text{if } c>a+b \text{ and } l_1'<l_3'$$

$$=\frac{1}{\sqrt{2\pi}}\frac{1}{ab}\left(\frac{ab}{c}\right)^{l_3'+1/2}(\sin\varphi_c)^{l_3'}P_{l_1'}^{-l_3'}(\cos\varphi_c) \quad \text{if } |a-b|<c<a+b.$$

N.B. If $|a-b|<c<a+b$, $c^2=a^2+b^2-2ab\cos\varphi_c$; if $c>a+b$, $c^2=a^2+b^2+2ab\,chU_c$.

However, the method fails to work when $n+l_1+l_2+l_3$ is odd. This case corresponds to a case of factorization $\lambda=\nu+2$, $\mu=-\rho$ and the F_4 function reduces to an F_1 function [4] with a positive integer index that yields a logarithmic singularity while for even parity this index is a negative integer and the corresponding F_1 is thus a polynomial. For these reasons, we give another method that is specific to the spherical Bessel functions.

3. Second method (spherical Bessel functions only). The method holds for any positive (or zero) integers l_1, l_2, l_3 $(-(l_1+l_2+l_3)\leqq n\leqq 2)$ and any real positive parameters a, b, c and it is specific to half-integer Bessel functions. Contrary to the preceding section, the final expression is not written as a sum of Legendre functions changing when going from a triangle to a nontriangle configuration: it appears as a finite sum of simple rational functions of a, b, c directly with a singularity that depends on the parity of the quantity $n+l_1+l_2+l_3$.

Integrals

$$A_{l_1,l_2,l_3,n}=\int_0^\infty t^n j_{l_1}(at)j_{l_2}(bt)j_{l_3}(ct)\,dt \tag{3.1}$$

may be calculated directly, using the expansion of the spherical Bessel functions. Following [1] we write

$$j_l(\rho)=\frac{e^{i\rho}}{2\rho}\sum_{m=0}^{l}\frac{\exp\left(i\pi/2(l-m+1)\right)}{m!}\frac{(l+m)!}{(l-m)!}\left(\frac{1}{2\rho}\right)^m+\text{c.c.} \tag{3.2}$$

where c.c. denotes the complex conjugate. The integrand is rewritten as follows:

$$\sum_{\{m_i\}} \sum_{\pm} \sum_{i} \prod \frac{(l_i+m_i)!}{m_i!(l_i-m_i)!} \exp\left(\mp \frac{i\pi}{2}(l_i-m_i+1)\right)$$

$$\cdot \frac{e^{it(\pm a\pm b\pm c)}}{(2a)^{m_1+1}(2b)^{m_2+1}(2c)^{m_3+1}} t^{n-3-m_1-m_2-m_3} \tag{3.3}$$

$$0 \leqq m_i \leqq l_i, \qquad i=1,2,3.$$

The integral

$$\int_0^\infty e^{it(\pm a\pm b\pm c)} t^{n-3-m_1-m_2-m_3}\, dt = \lim_{\varepsilon\to 0^+} \int_0^\infty e^{it(\pm a\pm b\pm c)}(t+\varepsilon)^{n-3-m_1-m_2-m_3}\, dt \tag{3.4}$$

is divergent, but the whole sum (3.3) is convergent. Then, we first calculate $\int_0^\infty e^{it(\pm a\pm b\pm c)}(t+\varepsilon)^{n-3-m_1-m_2-m_3}$, which exists for $\varepsilon>0$ and retain only its finite part when $\varepsilon\to 0^+$, *as it gives the only contribution to the final result.* The basic integral

$$I_p = \int_0^\infty \frac{e^{i\lambda t}}{(t+\varepsilon)^p}\, dt \tag{3.5}$$

$\lambda = \pm a \pm b \pm c$, real, is a polynomial in $1/\varepsilon$ plus a singular part I_1:

$$I_p = \frac{(i\lambda)^{p-1}}{(p-1)!} I_1 + \frac{1}{(p-1)!} \sum_{l=0}^{p-2} \frac{(p-2-l)!(i\lambda)^p}{\varepsilon^{p-1-l}}$$

and when $\varepsilon\to 0^+$, I_1 has a logarithmic behavior

$$I_1 = \int_0^\infty \frac{e^{i\lambda t}}{t+\varepsilon}\, dt = -\ln|\lambda| - \ln(e^\gamma\cdot\varepsilon) + \frac{i\pi}{2}\operatorname{sgn}(\lambda) + O(\varepsilon) \tag{3.6}$$

where γ is the Euler constant and $\operatorname{sgn}(x) = x/|x|$ is the sign function.

From (3.5)-(3.6), we get

$$\text{finite part } I_p = -\frac{(i\lambda)^{p-1}}{(p-1)!}\left[\ln|\lambda| + \ln e^\gamma - \frac{i\pi}{2}\operatorname{sgn}(\lambda)\right] \tag{3.7}$$

but $\ln(e^\gamma)$ will not appear in the final result as it cancels together with the $\ln\varepsilon$ terms. Using (3.7) and (3.3)-(3.4), we get all the integrals (3.1).

For example, we calculate integrals that escape the method of § 2.3:

$$\int_0^\infty j_0(at)j_0(bt)j_0(ct)\, dt = \frac{\pi}{2\sup(a,b,c)} \text{ outside the triangle}$$

$$= \frac{\pi}{8abc}\Lambda(\sqrt{a},\sqrt{b},\sqrt{c}), \qquad |a-b|<c<a+b$$

where $\Lambda(\sqrt{x},\sqrt{y},\sqrt{z}) = x^2+y^2+z^2-2xy-2yz-2zx$.

$$\int_0^\infty j_1(at)j_1(bt)j_0(ct)\, dt = 0 \quad \text{if } c>a+b$$

$$= \frac{\pi}{96a^2b^2c}[c-(a+b)]^2\{c^2+2c(a+b)-3(a-b)^2\},$$

$$|a-b|<c<a+b$$

$$= \frac{\pi}{3!}\frac{\inf(a,b)}{[\sup(a,b)]^2}, \qquad c<|a-b|$$

to be checked with the result in [1].

Both integrals are for $n+l_1+l_2+l_3$ even but $l_3+l_1-l_2+n-2$ negative. Note that a nonzero contribution appears outside the triangle. For odd $n+l_1+l_2+l_3$ we calculate

$$4abc\int_0^\infty tj_0(at)j_0(bt)j_0(ct)\,dt$$
$$= c\ln\left|\frac{c^2-(a-b)^2}{c^2-(a+b)^2}\right| + a\ln\left|\frac{a^2-(b-c)^2}{a^2-(b+c)^2}\right| + b\ln\left|\frac{b^2-(c-a)^2}{b^2-(c+a)^2}\right|,$$

which is symmetrical in a, b, c.

More generally, using expansion (3.3), we get for integrals (3.1) the final result:

$$\begin{aligned} A_{l_1l_2l_3n} = (-)^q\pi[&F_{l_1l_2l_3n}(a,b,c)\ \mathrm{sgn}\,(c+a+b)\\ &+(-)^{l_2}F_{l_1l_2l_3n}(a,-b,c)\ \mathrm{sgn}\,(c+a-b)\\ &+(-)^{l_1}F_{l_1l_2l_3n}(-a,b,c)\ \mathrm{sgn}\,(c-a+b)\\ &+(-)^{l_1+l_2}F_{l_1l_2l_3n}(-a,-b,c)\ \mathrm{sgn}\,(c-a-b)] \end{aligned} \tag{3.8}$$

when $l_1+l_2+l_3+n=2q$ is even ($\mathrm{sgn}\,(x)$ is again the sign function), and

$$\begin{aligned} A_{l_1l_2l_3n} = (-)^{q'+1}2[&F_{l_1l_2l_3n}(a,b,c)\ln|c+a+b|\\ &+(-)^{l_2}F_{l_1l_2l_3n}(a,-b,c)\ln|c+a-b|\\ &+(-)^{l_1}F_{l_1l_2l_3n}(-a,b,c)\ln|c-a+b|\\ &+(-)^{l_1+l_2}F_{l_1l_2l_3n}(-a,-b,+c)\ln|c-(a+b)|] \end{aligned} \tag{3.9}$$

when $l_1+l_2+l_3+n=2q'-1$ is odd. The function F includes all the combinatory coefficients of expansion (3.3):

$$\begin{aligned} F_{l_1l_2l_3n}(a,b,c) = \sum_{\{m_i\}} &\frac{(-)^{m_1+m_2+m_3}}{(2a)^{m_1+1}(2b)^{m_2+1}(2c)^{m_3+1}}\prod_{i=1}^{3}\frac{(l_i+m_i)!}{(l_i-m_i)!\,m_i!}\\ &\times\frac{(c+a+b)^{m_1+m_2+m_3+2-n}}{(m_1+m_2+m_3+2-n)!} \end{aligned} \tag{3.10}$$

TABLE 2

Expression of the integrals with three spherical Bessel functions in terms of rational functions of a, b, c and of the singularity.

$$\int_0^\infty t^n j_{l_1}(at)j_{l_2}(bt)j_{l_3}(ct)\,dt$$

$$= (-)^q\pi[F_{l_1l_2l_3n}(a,b,c)\ \mathrm{sgn}\,(c+a+b)+(-)^{l_2}F_{l_1l_2l_3n}(a,-b,c)\ \mathrm{sgn}\,(c+a-b)$$
$$+(-)^{l_1}F_{l_1l_2l_3n}(-a,b,c)\ \mathrm{sgn}\,(c-a+b)+(-)^{l_1+l_2}F_{l_1l_2l_3n}(-a,-b,c)\ \mathrm{sgn}\,(c-a-b)]$$
$$\text{if } n+l_1+l_2+l_3=2q$$

$$= (-)^{q'+1}2[F_{l_1l_2l_3n}(a,b,c)\ln|c+a+b|$$
$$+(-)^{l_2}F_{l_1l_2l_3n}(a,-b,c)\ln|c+a-b|+(-)^{l_1}F_{l_1l_2l_3}(a,b,c)\ln|c-a+b|$$
$$+(-)^{l_1+l_2}F_{l_1l_2l_3n}(-a,-b,c)\ln|c-a-b|]$$
$$\text{if } n+l_1+l_2+l_3=2q'-1$$

with

$$\mathrm{sgn}\,(x)=\frac{x}{|x|},$$

$$F_{l_1l_2l_3n}(a,b,c)=\sum_{\{m_i\}}\frac{(-)^{m_1+m_2+m_3}}{(2a)^{m_1+1}(2b)^{m_2+1}(2c)^{m_3+1}}\prod_{i=1}^{3}\frac{(l_i+m_i)!}{(l_i-m_i)!\,m_i!}\,\frac{(c+a+b)^{m_1+m_2+m_3+2-n}}{(m_1+m_2+m_3+2-n)!}.$$

Results (3.8), (3.9) are not expressed in terms of special functions as in the previous section. Nevertheless, they are not too difficult to handle for practical purposes, both because functions $F_{l_1 l_2 l_3 n}$ are simple rational functions and because the different behavior when a, b, c form or do not form a triangle appears through a unique singular multiplicative factor, i.e., the logarithmic term when $n+l_1+l_2+l_3$ is odd, the sign function when it is even.

Note once again that it is by no means necessary to assume that $n+l_1+l_2+l_3$ is even, nor that $n \geqq 0$ as the proof holds for any integer n such that $2-n \geqq 0$ and $n+l_1+l_2+l_3 \geqq 0$. The results of this section are summarized in Table 2.

Appendix A. Solution when $n=0$, $l_1=0$, $l_2=l_3=l$. We rewrite, using integration by parts,

$$I=\int_0^\infty t^{-3/2} J_{1/2}(at) J_\nu(bt) J_\nu(ct)\, dt \qquad (\nu = l+\tfrac{1}{2})$$

as

$$\begin{aligned} 2\nu I &= a\int_0^\infty t^{-1/2} J_{3/2}(at) J_\nu(bt) J_\nu(ct)\, dt \\ &\quad + \int_0^\infty t^{-1/2} J_{1/2}(at)[bJ_{\nu-1}(bt)J_\nu(ct) + cJ_\nu(bt)J_{\nu-1}(ct)]\, dt \\ &= -a \operatorname{Re} \mathcal{J}^{0\ 0}_{\nu\ 3/2}(b,c,a) + b \operatorname{Re} \mathcal{J}^{\nu-1-1/2\ 0}_{1/2\ \nu}(a,b,c) \\ &\quad + c \operatorname{Re} \mathcal{J}^{\nu-1-1/2\ 0}_{1/2\ \nu}(a,c,b). \end{aligned}$$

Thus

$$(2l+1)I = -a \operatorname{Re} \mathcal{J}^{0\ 0}_{l+1/2\ 3/2}(b,c,a) + b \operatorname{Re} \mathcal{J}^{l-1\ 0}_{1/2\ l+1/2}(a,b,c) + \operatorname{Re} c\mathcal{J}^{l-1\ 0}_{1/2\ l+1/2}(a,c,b)$$

where the dependence on parameters a, b, c is made explicit. Now, from (2.6) and (2.13), we get easily the (known) result when a, b, c do not form a triangle [8]. We have

$$\int_0^\infty j_0(at) j_l(bt) j_l(ct)\, dt = \left(\frac{\pi}{2}\right)^{3/2} \frac{1}{\sqrt{abc}} I$$

(A1)

$$\begin{aligned} &= 0 && \text{if } a > (b+c) \\ &= (2l+1)^{-1} \frac{\pi}{2} \left(\frac{b}{c}\right)^l \frac{1}{c} && \text{if } c > a+b \\ &= (2l+1)^{-1} \frac{\pi}{2} \left(\frac{c}{b}\right)^l \frac{1}{b} && \text{if } b > a+c. \end{aligned}$$

In the triangle case $|a-b| < c < a+b$, the result is a more complicated sum, namely

$$\begin{aligned} (2l+1)I = \Bigg\{ & -a \operatorname{Re} \mathcal{J}^{0\ 0}_{l+1/2\ 3/2}(b,c,a) \\ & + \sum_{m=0}^{l-1} (-)^m C^m_{l-1} \left[\left(\frac{b}{c}\right)^{l-1-m} \left(\frac{a}{c}\right)^m b \operatorname{Re} \mathcal{J}^{0\ 0}_{m+1/2\ 3/2}(a,b,c) + b \leftrightarrow c \right] \Bigg\}. \end{aligned}$$

Now, from (2.6)

$$\operatorname{Re} \mathcal{J}^0_{m+1/2\ 3/2}{}^0(a, b, c) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{ab}{c}} \cdot \frac{1}{c} P_m^{-1}(\cos \varphi_c) \sin \varphi_c,$$

$$\operatorname{Re} \mathcal{J}^0_{l+1/2\ 3/2}{}^0(b, c, a) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{cb}{a}} \cdot \frac{1}{a} P_l^{-1}(\cos \varphi_a) \sin \varphi_a,$$

(A2)
$$\int_0^\infty j_0(at) j_l(bt) j_l(ct)\, dt = (2l+1)^{-1} \frac{\pi}{4} \left\{ -\frac{\sin \varphi_a}{a} P_l^{-1}(\cos \varphi_a) + \sum_{m=0}^{l-1} (-)^m C_{l-1}^m \times \left[\left(\frac{b}{c}\right)^{l-m} \left(\frac{a}{c}\right)^m \frac{\sin \varphi_c}{c} P_m^{-1}(\cos \varphi_c) + b \leftrightarrow c \right] \right\}.$$

Note the functional relation

$$\mathcal{J}^{l-1\ 0}_{3/2\ l+1/2}(a, b, c) \equiv \mathcal{J}^0_{l+1/2\ 3/2}{}^0(b, c, a)$$

where the sum of $l-1$ Legendre polynomials in variable $\cos \varphi_c$ is expressed as a unique Legendre polynomial in $\cos \varphi_a$.

Appendix B. The proof goes as follows:

$$\mathcal{Q}_l^m(Z) = (Z^2-1)^{m/2} \frac{d^m}{dZ^m} \mathcal{Q}_l(Z)$$

with

$$\mathcal{Q}_l(Z) = \left\{ \frac{1}{2} \mathcal{P}_l(Z) \ln \frac{Z+1}{Z-1} + W_{l-1}(Z) \right\}$$

where $\mathcal{P}_l(Z)$ is the usual Legendre polynomial and $W_{l-1}(Z)$ is a polynomial of degree $l-1$ in Z. Thus

$$\text{for } m > l \quad \frac{d^m}{dZ^m} W_{l-1}(Z) = \frac{d^m}{dZ^m} \mathcal{P}_l(Z) = 0,$$

which implies that there is no logarithmic term in $\mathcal{Q}_l^m(Z)$. Furthermore,

$$\frac{d^p}{dZ^p} \ln \left\{ \frac{(Z+1)}{(Z-1)} \right\} = \frac{V_{p-1}(Z)}{(Z^2-1)^p}$$

where $V_{p-1}(Z)$ is a polynomial of degree $p-1$. The Leibniz formula yields

$$(Z^2-1)^{m/2} \mathcal{Q}_l^m(Z) = \frac{1}{2} (Z^2-1)^m \sum_{p=0}^{l} \frac{d^p \mathcal{P}_l(Z)}{dZ^p} \times \frac{V_{m-p-1}(Z)}{(Z^2-1)^{m-p}} C_m^p$$

$$= \frac{1}{2} \sum_{p=0}^{l} \frac{d^p \mathcal{P}_l(Z)}{dZ^p} \times V_{m-p-1}(Z) C_m^p (Z^2-1)^p$$

$$= \frac{1}{2} \sum_{p=0}^{l} (Z^2-1)^{p/2} \mathcal{P}_l^p(Z) \times V_{m-p-1}(Z) C_m^p.$$

Since $(Z^2-1)^{p/2} \mathcal{P}_l^p(Z)$ is a polynomial of degree $l+p$, $(Z^2-1)^{m/2} \mathcal{Q}_l^m(Z)$ is a polynomial of degree $l+m-1$.

REFERENCES

[1] A. BAEZA, B. BILWES, R. BILWES, J. DIAZ, J. L. FERRERO, AND J. RAYNAL, Nuclear Phys A, 437 (1985), pp. 93-116.

[2] A. D. JACKSON AND L. C. MAXIMON, *Integrals of products of Bessel functions*, SIAM J. Math. Anal., 3 (1972), pp. 446-460.

[3] W. N. BAILEY, Proc. London Math. Soc. (3), 40 (1936), pp. 37-48.

[4] P. APPELL AND J. KAMPÉ DE FERIET, *Fonctions hypergéométriques et hypersphériques*, Gauthier Villars, Paris, 1926.

[5a] W. N. BAILEY, Quart J. Math., 4 (1933), pp. 305-308.

[5b] ———, Quart. J. Math., 5 (1934), pp. 291-292.

[6a] A. GERVOIS AND H. NAVELET, *Integrals of three Bessel functions and Legendre functions*, I, J. Math. Phys., 26 (1985), pp. 633-644.

[6b] ———, *Integrals of three Bessel functions and Legendre functions*, II, J. Math. Phys., 26 (1985), pp. 645-655.

[7a] ———, *Some integrals involving three modified Bessel functions*, I, J. Math. Phys., 27 (1986), pp. 682-687.

[7b] ———, *Some integrals involving three modified Bessel functions*, II, J. Math. Phys., 27 (1986), pp. 688-695.

[8] A. ERDELYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Tables of Integral Transforms*, McGraw-Hill, New York, 1953.

[9] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals Series and Products*, Academic Press, New York, 1965.

[10] A. GERVOIS AND H. NAVELET, *Some integrals involving three modified Bessel functions when their arguments satisfy the triangle inequalities*, J. Math. Phys., 25 (1984), pp. 3350-3356.

[11] W. MAGNUS, F. OBERHETTINGER, AND R. P. SONI, *Formulas and Theorems for Special Functions of Mathematical Physics*, Springer-Verlag, Berlin, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 20, No. 4, pp. 1019–1034, July 1989

015

q-EXTENSIONS OF CLAUSEN'S FORMULA AND OF THE INEQUALITIES USED BY DE BRANGES IN HIS PROOF OF THE BIEBERBACH, ROBERTSON, AND MILIN CONJECTURES*

GEORGE GASPER†

Abstract. A q-extension of the terminating form of Clausen's ${}_3F_2$ series representation for the square of a ${}_2F_1(a,b;a+b+1/2;z)$ series is derived. It is used to prove the nonnegativity of certain basic hypergeometric series and to derive q-extensions of the inequalities and differential equations de Branges used in his proof of the Bieberbach, Robertson, and Milin conjectures.

Key words. Clausen's formula, basic hypergeometric series, nonnegative polynomials, inequalities, q-difference equations, Bieberbach, Robertson, and Milin conjectures

AMS(MOS) subject classifications. primary 33A99; secondary 30C50

1. Introduction.

In 1828 Clausen [15] used second- and third-order differential equations to prove the formula

$$\left\{{}_2F_1\left[\begin{matrix}a,b\\a+b+1/2\end{matrix};z\right]\right\}^2={}_3F_2\left[\begin{matrix}2a,2b,a+b\\2a+2b,a+b+1/2\end{matrix};z\right],\qquad |z|<1. \tag{1.1}$$

The above ${}_2F_1$ and ${}_3F_2$ are special cases of ${}_rF_s$ hypergeometric series defined by

$${}_rF_s\left[\begin{matrix}a_1,\cdots,a_r\\b_1,\cdots,b_s\end{matrix};z\right]=\sum_{n=0}^{\infty}\frac{(a_1)_n\cdots(a_r)_n}{n!(b_1)_n\cdots(b_s)_n}z^n,$$

where $(a)_n$ is the shifted factorial defined by

$$(a)_n=\prod_{k=0}^{n-1}(a+k).$$

Almost 150 years later, Clausen's formula was used in Askey and Gasper [4] to prove that

$${}_3F_2\left[\begin{matrix}-n,n+\alpha+2,(\alpha+1)/2\\\alpha+1,(\alpha+3)/2\end{matrix};\frac{1-x}{2}\right]\geq 0,\qquad -1\leq x\leq 1, \tag{1.2}$$

when $\alpha>-2$ and $n=0,1,2,\cdots$, and then this inequality was used to prove the positivity of certain important kernels involving sums of Jacobi polynomials (also see Askey [3, Lecture 8] and the extensions in Gasper [18], [19]). In 1984 the special cases $\alpha=2,4,6,\cdots$ of (1.2) were used by de Branges [11], [12] to complete the last part of his proof of the Milin [30, p. 55] conjecture that if f is in the class S of functions

$$f(z)=z+c_2z^2+c_3z^3+\cdots$$

that are analytic and univalent in the unit disk $|z|<1$ and if

$$\log\frac{f(z)}{z}=2\sum_{k=1}^{\infty}\gamma_kz^k,$$

*Received by the editors May 2, 1988; accepted for publication June 21, 1988. This work was supported in part by the National Science Foundation under grant DMS-8601901.

†Department of Mathematics, Northwestern University, Evanston, Illinois 60208.

then

$$\sum_{k=1}^{n} k(n+1-k)|\gamma_k|^2 \leq \sum_{k=1}^{n} \frac{n+1-k}{k}, \qquad n = 1, 2, \cdots. \tag{1.3}$$

It was already known that Milin's conjecture implied Robertson's [33] conjecture that if f is an odd function in S, then

$$\sum_{k=1}^{n} |c_{2k-1}|^2 \leq n, \qquad n = 2, 3, \cdots, \tag{1.4}$$

and that Robertson's conjecture implied Bieberbach's [9] conjecture that if f is in S, then

$$|c_n| \leq n, \qquad n = 2, 3, \cdots. \tag{1.5}$$

Since ${}_rF_s$ series are limit cases of ${}_r\phi_s$ basic hypergeometric series [25], [38]

$${}_r\phi_s\left[\begin{matrix} a_1, \cdots, a_r \\ b_1, \cdots, b_s \end{matrix}; q, z\right] = \sum_{n=0}^{\infty} \frac{(a_1, \cdots, a_r; q)_n}{(q, b_1, \cdots, b_s; q)_n} [(-1)^n q^{\binom{n}{2}}]^{1+s-r} z^n, \tag{1.6}$$

where $\binom{n}{2} = n(n-1)/2$,

$$(a_1, a_2, \cdots, a_r; q)_n = (a_1; q)_n (a_2; q)_n \cdots (a_r; q)_n$$

and $(a; q)_n$ is the q-shifted factorial defined by

$$(a; q)_n = \prod_{k=0}^{n-1} (1 - aq^k),$$

it is natural to search for q-extensions (also called q-analogues and, in the terminology of [10], quantum generalizations) of Clausen's formula (1.1), the inequalities (1.2), and of the other parts of de Branges' proof of the Milin conjecture.

In this paper we will derive a q-extension of Clausen's formula (1.1) for terminating series and various q-extensions of the inequalities (1.2)and of some other inequalities. In addition, since the existence of decreasing solutions of de Branges' differential equations

$$\sigma_n(t) + \frac{t}{n}\sigma_n'(t) = \sigma_{n+1}(t) - \frac{t}{n+1}\sigma_{n+1}'(t), \qquad 1 \leq t < \infty, \tag{1.7}$$

played a crucial role in his proof of the Milin conjecture, we derive q-extensions of (1.7) and show that they have solutions which have negative first q-derivatives. Some prospects for further research are pointed out.

2. q-Extensions of Clausen's formula (1.1). In 1940 Jackson [27] derived a general theorem about solutions of q^θ equations, where q^θ is the operator $\exp((\log q) \cdot x\frac{d}{dx})$, which gives [28, p. 171] the product formula

$$\begin{aligned} &{}_2\phi_1\left[\begin{matrix} q^{2a}, q^{2b} \\ q^{2a+2b+1} \end{matrix}; q^2, z\right] {}_2\phi_1\left[\begin{matrix} q^{2a}, q^{2b} \\ q^{2a+2b+1} \end{matrix}; q^2, qz\right] \\ &\quad = {}_4\phi_3\left[\begin{matrix} q^{2a}, q^{2b}, q^{a+b}, -q^{a+b} \\ q^{2a+2b}, q^{a+b+1/2}, -q^{a+b+1/2} \end{matrix}; q, z\right], \qquad |z| < 1, |q| < 1. \end{aligned} \tag{2.1}$$

Since

$$\lim_{q\uparrow 1}\frac{(q^a;q)_n}{(1-q)^n}=(a)_n,\qquad \lim_{q\uparrow 1}(-q^a;q)_n=2^n, \tag{2.2}$$

Clausen's formula (1.1) is a limit case of Jackson's product formula (2.1). However, unlike in (1.1), the left side of (2.1) is not a square and so (2.1) cannot be used to write sums of basic hypergeometric series as sums of squares of basic hypergeometric series as was done in [4], [18] for hypergeometric series to prove the nonnegativity of certain sums of hypergeometric series. Also, by considering negative integer values of a, we find that the series on the right side of (2.1) can assume negative values. It is natural to consider replacing the left side of (1.1) by

$$\left\{{}_2\phi_1\left[\begin{matrix}q^a,q^b\\q^{a+b+1/2}\end{matrix};q,z\right]\right\}^2 \tag{2.3}$$

but, unfortunately, this square of a ${}_2\phi_1$ series does not equal a basic hypergeometric series of the type in (1.6) as can be easily seen by computing the coefficient of z^2 in its power series expansion. Thus, in order to find a basic hypergeometric series which is the square of a basic hypergeometric series we are forced to look for another q-extension of (1.1).

One way to proceed is to recall that in [4] Clausen's formula was used to write (1.2) as a sum of squares of ultraspherical polynomials

$$\begin{aligned}C_n^\lambda(x)&=\frac{(2\lambda)_n}{n!}{}_2F_1\left[\begin{matrix}-n,n+2\lambda\\\lambda+1/2\end{matrix};\frac{1-x}{2}\right]\\&=\frac{(\lambda)_n}{n!}e^{in\theta}{}_2F_1\left[\begin{matrix}-n,\lambda\\1-n-\lambda\end{matrix};e^{-2i\theta}\right],\qquad x=\cos\theta,\end{aligned} \tag{2.4}$$

and to recall that in his work [34]–[36] during the 1890's on the now famous Rogers–Ramanujan identities, Rogers [36] considered the q-extension

$$C_n(x;\beta\mid q)=\frac{(\beta;q)_n}{(q;q)_n}e^{in\theta}{}_2\phi_1\left[\begin{matrix}q^{-n},\beta\\q^{1-n}\beta^{-1}\end{matrix};q,q\beta^{-1}e^{-2i\theta}\right],\quad x=\cos\theta, \tag{2.5}$$

of (2.4). Askey and Ismail [6] showed that these polynomials were orthogonal on $(-1,1)$ with respect to an absolutely continuous weight function and called them the continuous q-ultraspherical polynomials to distinguish them from the (discrete) q-ultraspherical polynomials

$$C_n^\lambda(x;q)=\frac{(q^{2\lambda};q)_n}{(q;q)_n}{}_2\phi_1\left[\begin{matrix}q^{-n},q^{n+2\lambda}\\q^{\lambda+1/2}\end{matrix};q,qx\right] \tag{2.6}$$

which are orthogonal [2, (3.8)] with respect to a discrete measure with point masses at $x=q^k$, $k=0,1,2,\cdots$. They also showed that

$$C_n^\lambda(x)=\lim_{q\uparrow 1}C_n(x;q^\lambda\mid q) \tag{2.7}$$

and

$$C_n(\cos\theta;\beta\mid q)=\frac{(\beta^2;q)_n}{\beta^{n/2}(q;q)_n}{}_4\phi_3\left[\begin{matrix}q^{-n},\beta^2q^n,\beta^{1/2}e^{i\theta},\beta^{1/2}e^{-i\theta}\\\beta q^{1/2},-\beta q^{1/2},-\beta\end{matrix};q,q\right]. \tag{2.8}$$

In 1895 Rogers [36, p. 29] employed an induction argument to prove the linearization formula

$$C_m(x;\beta \mid q)C_n(x;\beta \mid q) = \sum_{k=0}^{\min(m,n)} \frac{(q;q)_{m+n-2k}(\beta;q)_{m-k}(\beta;q)_{n-k}}{(q;q)_k(q;q)_{m-k}(q;q)_{n-k}} \cdot \frac{(\beta;q)_k(\beta^2;q)_{m+n-k}(1-\beta q^{m+n-2k})}{(\beta q;q)_{m+n-k}(\beta^2;q)_{m+n-2k}(1-\beta)} C_{m+n-2k}(x;\beta \mid q). \tag{2.9}$$

A simple computational proof of (2.9) was given by the author in [20]. For additional proofs, see Bressoud [14] and Rahman [31]. Note that if we use (2.5) or (2.8) on the right side of (2.9), we get a double sum that does not reduce to a single sum when $n = m$, even after changing the order of summation and trying to apply known summation formulas. This is also true when (2.8) is replaced by the formulas obtained by applying Sears' transformation formula [37, (8.3)]

$${}_4\phi_3\left[\begin{matrix} a, b, c, q^{-n} \\ d, e, f \end{matrix}; q, q\right] = \frac{(de/ab, df/ab; q)_n}{(e, f; q)_n} \left(\frac{ab}{d}\right)^n {}_4\phi_3\left[\begin{matrix} d/a, d/b, c, q^{-n} \\ d, de/ab, df/ab \end{matrix}; q, q\right] \tag{2.10}$$

where $\text{def} = abcq^{1-n}$ and $n = 0, 1, 2, \cdots$, to the ${}_4\phi_3$ series in (2.8).

The right side of (2.8) suggests that we should still look at expansions involving $e^{i\theta}$ and $e^{-i\theta}$ among the parameters. Since the polynomials on the right side of (2.9) are of even degree in x when $n = m$, we could try to use the expansion [7, p. 41]

$$C_{2n}(\cos\theta;\beta \mid q) = \frac{(\beta^2;q)_{2n}(-q, -q^{1/2};q)_n}{(q;q)_{2n}(-\beta, -\beta q^{1/2};q)_n} q^{-n/2} \cdot {}_4\phi_3\left[\begin{matrix} q^{-n}, \beta q^n, q^{1/2}e^{2i\theta}, q^{1/2}e^{-2i\theta} \\ \beta q^{1/2}, -q^{1/2}, -q \end{matrix}; q, q\right], \tag{2.11}$$

in which the ${}_4\phi_3$ series terminates after $n+1$ terms, even though $C_{2n}(x;\beta|q)$ is a polynomial of degree $2n$ in x. But, by using (2.11) in the right side of (2.9) and changing the order of summation we get a sum of terminating very well poised ${}_8\phi_7$ series that are not balanced and hence not summable by Jackson's formula [38, (3.3.1.1)].

However, if we apply (2.10) to (2.11) to get

$$C_{2n}(\cos\theta;\beta \mid q) = \frac{(\beta^2;q)_{2n}}{\beta^n(q;q)_{2n}} {}_4\phi_3\left[\begin{matrix} q^{-n}, \beta q^n, \beta e^{2i\theta}, \beta e^{-2i\theta} \\ \beta q^{1/2}, -\beta q^{1/2}, -\beta \end{matrix}; q, q\right] \tag{2.12}$$

and use (2.12) in the right side of (2.9) we obtain

$$\begin{aligned} &\{C_n(\cos\theta;\beta \mid q)\}^2 \\ &\quad = \sum_{k=0}^{n} \frac{(\beta, \beta; q)_{n-k}(\beta;q)_k(\beta^2;q)_{2n-k}(1-\beta q^{2n-2k})}{(q, q; q)_{n-k}(q;q)_k(\beta q;q)_{2n-k}(1-\beta)} \\ &\quad\quad \cdot \beta^{k-n} \sum_{j=0}^{n-k} \frac{(q^{k-n}, \beta q^{n-k}, \beta e^{2i\theta}, \beta e^{-2i\theta}; q)_j}{(q, \beta q^{1/2}, -\beta q^{1/2}, -\beta; q)_j} q^j \\ &\quad = \frac{(\beta^2;q)_{2n}(\beta, \beta; q)_n}{\beta^n(\beta;q)_{2n}(q, q; q)_n} \sum_{j=0}^{n} \frac{(q^{-n}, \beta q^n, \beta e^{2i\theta}, \beta e^{-2i\theta}; q)_j}{(q, \beta q^{1/2}, -\beta q^{1/2}, -\beta; q)_j} q^j \\ &\quad\quad \cdot {}_6\phi_5\left[\begin{matrix} \beta^{-1}q^{-2n}, \beta^{-1/2}q^{1-n}, -\beta^{-1/2}q^{1-n}, \beta, q^{j-n}, q^{-n} \\ \beta^{-1/2}q^{-n}, -\beta^{-1/2}q^{-n}, \beta^{-2}q^{1-2n}, \beta^{-1}q^{1-n-j}, \beta^{-1}q^{1-n} \end{matrix}; q, \beta^{-2}q^{1-j}\right] \end{aligned} \tag{2.13}$$

in which, fortunately, the ${}_6\phi_5$ series is summable by the summation formula [38, (3.3.1.4)]

$$(2.14)\qquad {}_6\phi_5\left[\begin{matrix}a, qa^{1/2}, -qa^{1/2}, b, c, q^{-n}\\ a^{1/2}, -a^{1/2}, aq/b, aq/c, aq^{n+1}\end{matrix}; q, \frac{aq^{n+1}}{bc}\right] = \frac{(aq, aq/bc; q)_n}{(aq/b, aq/c; q)_n}, \qquad n = 0, 1, 2, \cdots,$$

where four misprints have been corrected. Using (2.14) to sum the ${}_6\phi_5$ series in (2.13) gives

$$(2.15)\qquad \{C_n(\cos\theta; \beta \mid q)\}^2 = \frac{(\beta^2, \beta^2; q)_n}{(q, q; q)_n}\beta^{-n}{}_5\phi_4\left[\begin{matrix}q^{-n}, \beta^2 q^n, \beta, \beta e^{2i\theta}, \beta e^{-2i\theta}\\ \beta^2, \beta q^{1/2}, -\beta q^{1/2}, -\beta\end{matrix}; q, q\right]$$

and hence, by (2.8), we have the following q-extension of the terminating case of Clausen's formula (1.1)

(2.16)

$$\left\{{}_4\phi_3\left[\begin{matrix}q^{-n}, \beta^2 q^n, \beta^{1/2}e^{i\theta}, \beta^{1/2}e^{-i\theta}\\ \beta q^{1/2}, -\beta q^{1/2}, -\beta\end{matrix}; q, q\right]\right\}^2 = {}_5\phi_4\left[\begin{matrix}q^{-n}, \beta^2 q^n, \beta, \beta e^{2i\theta}, \beta e^{-2i\theta}\\ \beta^2, \beta q^{1/2}, -\beta q^{1/2}, -\beta\end{matrix}; q, q\right],$$

where $n = 0, 1, 2, \cdots$. This formula was derived independently by Mizan Rahman. He and the author independently observed that it can be derived by using (2.8) inside the Rahman and Verma [32, (1.20)] integral representation for the product of two continuous q-ultraspherical polynomials and then integrating termwise to get (2.15) and hence (2.16).

By setting $a = \beta^{1/2}e^{i\theta}$, $b = \beta^{1/2}e^{-i\theta}$ and $z = \beta q^n$, formula (2.16) can be written in the form

(2.17)

$$\left\{{}_4\phi_3\left[\begin{matrix}a, b, abz, ab/z\\ abq^{1/2}, -abq^{1/2}, -ab\end{matrix}; q, q\right]\right\}^2 = {}_5\phi_4\left[\begin{matrix}a^2, b^2, ab, abz, ab/z\\ a^2b^2, abq^{1/2}, -abq^{1/2}, -ab\end{matrix}; q, q\right],$$

which holds when the series on both sides terminate. For, by (2.16), (2.17) holds when abz or ab/z is a negative integer power of q and, if a or b is a negative integer power of q, then both sides of (2.17) are rational functions of z which are equal for $z = abq^n$, $n = 0, 1, 2, \cdots$, and hence must be equal for all (complex) values of z. Notice that by replacing a, b, z in (2.17) by q^a, q^b, $e^{i\theta}$ with $x = \cos\theta$ and letting $q \uparrow 1$, we get Clausen's formula (1.1) with $z = (1 - x)/2$ for the terminating case when a or b is a negative integer.

To see that (2.17) does not hold in the nonterminating case, it suffices to observe that if, e.g., (2.17) held for $b = 0$, then it would follow fromthe q-binomial theorem [38, (3.2.2.11)] that $((aq; q)_\infty)^2 = (a^2q; q)_\infty(q; q)_\infty$, which is clearly false for, e.g., $a = q^{-1/2}$. A nonterminating q-extension of (1.1) containing the square of an ${}_8\phi_7$ series and the sum of two ${}_5\phi_4$ series will be given in [26].

Note that, in addition to the formulas that follow when (2.10) is applied to the ${}_4\phi_3$ in (2.17), we can apply the quadratic transformation formula [7, (3.2)]

$$(2.18)\qquad {}_4\phi_3\left[\begin{matrix}a^2, b^2q, c, d\\ abq, -abq, -cd\end{matrix}; q, q\right] = {}_4\phi_3\left[\begin{matrix}a^2, b^2q, c^2, d^2\\ a^2b^2q^2, -cd, -cdq\end{matrix}; q^2, q^2\right],$$

which holds when both series terminate, to obtain that (2.17) is equivalent to the formula

$$(2.19)\qquad \left\{{}_4\phi_3\left[\begin{matrix}a^2, b^2, abz, ab/z\\ a^2b^2q, -ab, -abq\end{matrix}; q^2, q^2\right]\right\}^2 = {}_5\phi_4\left[\begin{matrix}a^2, b^2, ab, abz, ab/z\\ a^2b^2, abq^{1/2}, -abq^{1/2}, -ab\end{matrix}; q, q\right]$$

when both series terminate. Also, if we replace a, b, z, q in (2.17) by their squares and apply (2.18), we obtain that

(2.20)
$$\left\{{}_4\phi_3\left[\begin{matrix}a^2, b^2, abz, ab/z\\ abq^{1/2}, -abq^{1/2}, -a^2b^2\end{matrix}; q, q\right]\right\}^2 = {}_5\phi_4\left[\begin{matrix}a^4, b^4, a^2b^2, a^2b^2z^2, a^2b^2/z^2\\ a^4b^4, a^2b^2q, -a^2b^2q, -a^2b^2\end{matrix}; q^2, q^2\right]$$

when both series terminate.

Another proof of (2.17) can be given by observing that from the product formulas (2.8) or (2.10) in Gasper and Rahman [24] it follows that if a, b, abz, or ab/z is a negative integer power of q, then

(2.21)
$$\begin{aligned}
&\left\{{}_4\phi_3\left[\begin{matrix}abz, ab/z, a, b\\ -ab, abq^{1/2}, -abq^{1/2}\end{matrix}; q, q\right]\right\}^2\\
&= \sum_{r\geq 0}\sum_{s\geq 0}\frac{(abz, ab/z; q)_{r+s}(a, b, -a, -b; q)_r}{(-ab, -ab; q)_{r+s}(q, -q, abq^{1/2}, -abq^{1/2}; q)_r}\\
&\quad\cdot\frac{(a, b, -a, -b; q)_s}{(q, -1, abq^{1/2}, -abq^{1/2}; q)_s}\,\frac{1+q^{r-s}}{1+q^s}\,q^{r+s}\\
&= \sum_{k\geq 0}\frac{(abz, ab/z, a, -a, b, -b; q)_k}{(q, -1, abq^{1/2}, -abq^{1/2}, -ab, -ab; q)_k}\,q^k\\
&\quad\cdot\sum_{s=0}^{k}\frac{(q^{-k}, -q^{-k}, a, -a, b, -b, a^{-1}b^{-1}q^{1/2-k}, -a^{-1}b^{-1}q^{1/2-k}; q)_s}{(q, -q, a^{-1}q^{1-k}, -a^{-1}q^{1-k}, b^{-1}q^{1-k}, -b^{-1}q^{1-k}, abq^{1/2}, -abq^{1/2}; q)_s}\,\frac{1+q^{2s-k}}{1+q^{-k}}\,q^{2s}\\
&= \sum_{k\geq 0}\frac{(abz, ab/z, a, b, -a, -b; q)_k}{(q, -1, abq^{1/2}, -abq^{1/2}, -ab, -ab; q)_k}\,q^k\\
&\quad\cdot{}_5\phi_4\left[\begin{matrix}q^{-2k}, -q^{2-k}, a^2, b^2, a^{-2}b^{-2}q^{1-2k}\\ -q^{-k}, a^{-2}q^{2-2k}, b^{-2}q^{2-2k}, a^2b^2q\end{matrix}; q^2, q^2\right]\\
&= {}_5\phi_4\left[\begin{matrix}a^2, b^2, ab, abz, ab/z\\ a^2b^2, abq^{1/2}, abq^{1/2}, -ab\end{matrix}; q, q\right],
\end{aligned}$$

since it can be shown that

(2.22)
$$\begin{aligned}
&{}_5\phi_4\left[q^{-2k}, -q^{2-k}, a^2, b^2, a^{-2}b^{-2}q^{1-2k} - q^{-k}, a^{-2}q^{2-2k}, b^{-2}q^{2-2k}, a^2b^2q; q^2, q^2\right]\\
&\qquad = \frac{(-1, a^2, b^2, ab, -ab; q)_k}{(a^2b^2, a, -a, b, -b; q)_k}, \qquad k = 0, 1, 2, \cdots,
\end{aligned}$$

by using the case $d = eq^{1/2} = (aq)^{1/2}$ of Jackson's summation formula [38, (3.3.1.1)]. A slightly more direct proof of (2.17) can be given by starting with the expansion [24, (2.2)] used to derive [24, (2.8)].

3. q-Extensions of (1.2). At the last step in his proof of the Milin conjecture, de Branges [12] used the fact that for any positive integer r the functions

(3.1)
$$\sigma_n(t) = \frac{n\Gamma(n+r+2)}{\Gamma(2n+2)\Gamma(r+1-n)}\int_t^\infty {}_3F_2\left[\begin{matrix}n-r, n+r+2, n+1/2\\ 2n+1, n+3/2\end{matrix}; s^{-1}\right]s^{-n-1}ds$$

when $n = 1, \cdots, r$ and $\sigma_n(t) = 0$ when $n > r$, satisfy the differential equations (1.7) and

$$(3.2)\quad \sigma_n'(t)) = -\frac{n\Gamma(n+r+2)t^{-n-1}}{\Gamma(2n+2)\Gamma(r+1-n))}{}_3F_2\left[\begin{matrix} n-r, n+r+2, n+1/2 \\ 2n+1, n+3/2 \end{matrix}; t^{-1}\right] \le 0$$

for $t \ge 1$ when $n = 1, \cdots, r$. In [13], to estimate the coefficients of powers of unbounded Riemann mapping functions, he used the fact that the more general functions

(3.3)

$$\sigma_n(t) = \frac{\Gamma(n+1)\Gamma(n+r+2\nu+2\lambda+1)4^{-n}t^{2\nu}}{\Gamma(n+\nu+1)\Gamma(n+2\nu)\Gamma(n+\nu+\lambda+1)\Gamma(r+1-n)} \cdot \int_t^\infty {}_3F_2\left[\begin{matrix} n-r, n+r+2\nu+2\lambda+1, n+\nu+1/2 \\ 2n+2\nu+1, n+\nu+\lambda+1 \end{matrix}; s^{-1}\right] s^{-n-2\nu-1}ds$$

when $n = 1, \cdots, r$ and $\sigma_n(t) = 0$ when $n > r$, satisfy the differential equations

$$(3.4)\quad \frac{n}{n+2\nu}\sigma_n(t) + \frac{t\sigma_n'(t)}{n+2\nu} = \frac{n+2\nu+1}{n+1}\sigma_{n+1}(t) - \frac{t\sigma_{n+1}'(t)}{n+1}$$

and

(3.5)

$$\frac{d}{dt}[t^{-2\nu}\sigma_n(t)] = -\frac{\Gamma(n+1)\Gamma(n+r+2\nu+2\lambda+1)4^{-n}}{\Gamma(n+\nu+1)\Gamma(n+2\nu)\Gamma(n+\nu+\lambda+1)\Gamma(r+1-n)} \cdot {}_3F_2\left[\begin{matrix} n-r, n+r+2\nu+2\lambda+1, n+\nu+1/2 \\ 2n+2\nu+1, n+\nu+\lambda+1 \end{matrix}; t^{-1}\right] \le 0$$

for $t \ge 1$ when $\nu > -1/2$, $\lambda \ge 0$, and $n = 1, \cdots, r$.

Since (3.5) reduces to (3.2) when $\nu = \lambda + 1/2 = 0$, in addition to deriving q-extensions of (1.2) we will derive q-extensions of the inequalities

$$(3.6)\quad {}_3F_2\left[\begin{matrix} -n, n+a, b \\ 2b, (a+1)/2 \end{matrix}; \frac{1-x}{2}\right] \ge 0, \qquad -1 \le x \le 1,$$

where $a \ge 2b > -1$ and $n = 0, 1, \cdots$, which imply the inequalities in (3.5) and reduce to (1.2) when $a = \alpha + 2$ and $b = (\alpha+1)/2$.

Let $0 < q < 1$, $n = 0, 1, 2, \cdots$, and let a, b, α, β, γ, δ, θ be real parameters. Then

$$(3.7)\quad \lim_{q\to 1} {}_5\phi_4\left[\begin{matrix} q^{-n}, q^{n+a}, q^b, q^\alpha e^{i\theta}, q^\beta e^{-i\theta} \\ q^{2b}, q^{(a+1)/2}, -q^\gamma, -q^\delta \end{matrix}; q, q\right] = {}_3F_2\left[\begin{matrix} -n, n+a, b \\ 2b, (a+1)/2 \end{matrix}; \frac{1-x}{2}\right], \quad x = \cos\theta,$$

and so, in order to derive a q-extension of (3.6), it suffices to find values of α, β, γ, δ for which the ${}_6\phi_5$ series in (3.7) are nonnegative when $a \ge 2b > -1$.

Observe that from (2.16)

$$(3.8)\quad {}_5\phi_4\left[\begin{matrix} q^{-n}, q^{n+2b}, q^b, q^b e^{i\theta}, q^b e^{-i\theta} \\ q^{2b}, q^{b+1/2}, -q^{b+1/2}, -q^b \end{matrix}; q, q\right] = \left\{{}_4\phi_3\left[\begin{matrix} q^{-n}, q^{n+2b}, q^{b/2}e^{i\theta/2}, q^{b/2}e^{-i\theta/2} \\ q^{b+1/2}, -q^{b+1/2}, -q^b \end{matrix}; q, q\right]\right\}^2 \ge 0,$$

which shows that the ${}_5\phi_4$ series in (3.7) are nonnegative when $a = 2b$, $\alpha = \beta = \delta = b$, and $\gamma = b + 1/2$. In view of the "sums of squares" method [18, §8], we will consider sums of the nonnegative ${}_5\phi_4$ series in (3.8).

The author showed in [18, §8] that besides the sum of squares of ultraspherical polynomials used in [4, (1.16)] to prove (1.2) we could also use the sum of squares in [18, (8.17)] and observed that these two expansions are special cases of the expansion [18, (8.18)]

$$
\begin{aligned}
&{}_3F_2\left[\begin{matrix}-n, n+\alpha+2, (\alpha+1)/2 \\ \alpha+1, (\alpha+3)/3\end{matrix}; (1-x^2)(1-y^2)\right] \\
&\quad = \sum_{j=0}^{n} \frac{n!(n+\alpha+2)_j \left(\frac{\alpha+2}{2}\right)_j}{j!(n-j)! \left(\frac{\alpha+3}{2}\right)_j (j+\alpha+1)_j} (1-y^2)^j \\
&\qquad \cdot \left\{ \frac{j!(n-j)!}{(\alpha+1)_j (2j+\alpha+2)_{n-j}} C_j^{(\alpha+1)/2}(x) C_{n-j}^{j+(\alpha+2)/2}(y) \right\}^2,
\end{aligned} \tag{3.9}
$$

which can be derived from the Fields and Wimp [17] expansion formula

$$
\begin{aligned}
{}_{r+t}F_{s+u}\left[\begin{matrix}a_R, c_T \\ b_S, d_U\end{matrix}; xw\right] &= \sum_{j=0}^{\infty} \frac{(a_R)_j (\alpha)_j (\beta)_j}{(b_S)_j (\gamma+j)_j} \frac{(-x)^j}{j!} \\
&\quad \cdot {}_{r+2}F_{s+1}\left[\begin{matrix}j+\alpha, j+\beta, j+a_R \\ 1+2j+\gamma, j+b_S\end{matrix}; x\right] {}_{t+2}F_{u+2}\left[\begin{matrix}-j, j+\gamma, c_T \\ \alpha, \beta, d_U\end{matrix}; w\right],
\end{aligned} \tag{3.10}
$$

where we used the contracted notation of representing $a_1, a_2, \cdots, a_r$ by a_R, $(a_1)_j (a_2)_j \cdots (a_r)_j$ by $(a_R)_j$, and $j+a_1, j+a_2, \cdots, j+a_r$ by $j+a_R$. Recently the author derived a bibasic extension [22, (4.5)] of (3.10) which contained Verma's [40] q-analogues and gave the general expansion [22, (4.7)]

(3.11)

$$
\begin{aligned}
&{}_{r+t}\phi_{s+u}\left[\begin{matrix}a_R, c_T \\ b_S, d_U\end{matrix}; q, xw\right] \\
&\quad = \sum_{j=0}^{\infty} \frac{(c_T, e_K, \sigma, \gamma q^{j+1}/\sigma; q)_j}{(q, d_U, f_M, \gamma q^j; q)_j} \left(\frac{x}{\sigma}\right)^j [(-1)^j q^{\binom{j}{2}}]^{u+m-t-k} \\
&\qquad \cdot {}_{t+k+4}\phi_{u+m+3}\left[\begin{matrix}\gamma q^{2j}/\sigma, q^{j+1}\sqrt{\gamma/\sigma}, -q^{j+1}\sqrt{\gamma/\sigma}, \sigma^{-1}, c_T q^j, e_K q^j \\ q^j\sqrt{\gamma/\sigma}, -q^j\sqrt{\gamma/\sigma}, \gamma q^{2j+1}, d_U q^j, f_M q^j\end{matrix}; q, xq^{j(u+m-t-k)}\right] \\
&\qquad \cdot {}_{r+m+2}\phi_{s+k+2}\left[\begin{matrix}q^{-j}, \gamma q^j, a_R, f_M \\ \gamma q^{j+1}/\sigma, q^{1-j}/\sigma, b_S, e_K\end{matrix}; q, wq\right],
\end{aligned} \tag{3.11}
$$

where we used a contracted notation analogous to that used in (3.10). Formulas (3.10) and (3.11) hold when the series terminate and when the parameters and variables are such that the series converge absolutely.

In this section we will use the following $\sigma \to \infty$ limit case of the $m = 2$, $f_1 = f_2 = 0$ case of (3.11)

$$
\begin{aligned}
&{}_{r+t}\phi_{s+u}\left[\begin{matrix} a_R, c_T \\ b_S, d_U \end{matrix}; q, xw\right] \\
&\quad = \sum_{j=0}^{\infty} \frac{(c_T, e_K; q)_j}{(q, d_U, \gamma q^j; q)_j} x^j [(-1)^j q^{\binom{j}{2}}]^{u+3-t-k} \\
&\quad \cdot {}_{t+k}\phi_{u+1}\left[\begin{matrix} c_T q^j, e_K q^j \\ \gamma q^{2j+1}, d_U q^j \end{matrix}; q, x q^{j(u+3-t-k)}\right] \\
&\quad \cdot {}_{r+2}\phi_{s+k}\left[\begin{matrix} q^{-j}, \gamma q^j, a_R \\ b_S, e_K \end{matrix}; q, wq\right],
\end{aligned} \tag{3.12}
$$

which is equivalent to [39, (3.1)]. Set

$$\gamma = q^{2b}, a_1 = q^b, a_2 = q^b e^{i\theta}, a_3 = q^b e^{-i\theta}, b_1 = q^{2b}, b_2 = -q^b,$$

$$c_1 = q^{-n}, c_2 = q^{n+a}, d_1 = q^{(a+1)/2} = -d_2, e_1 = q^{b+1/2} = -e_2, x = q, w = 1,$$

in the $r = 3$, $s = t = u = k = 2$ case of (3.12) to obtain

$$
\begin{aligned}
&{}_5\phi_4\left[\begin{matrix} q^{-n}, q^{n+a}, q^b, q^b e^{i\theta}, q^b e^{-i\theta} \\ q^{2b}, q^{(a+1)/2}, -q^{(a+1)/2}, -q^b \end{matrix}; q, q\right] \\
&\quad = \sum_{j=0}^{n} \frac{(q^{-n}, q^{n+a}, q^{b+1/2}, -q^{b+1/2}; q)_j}{(q, q^{(a+1)/2}, -q^{(a+1)/2}, q^{j+2b}; q)_j} (-1)^j q^{j+\binom{j}{2}} \\
&\quad \cdot {}_4\phi_3\left[\begin{matrix} q^{j-n}, q^{j+n+a}, q^{j+b+1/2}, -q^{j+b+1/2} \\ q^{2j+2b+1}, q^{j+(a+1)/2}, -q^{j+(a+1)/2} \end{matrix}; q, q\right] \\
&\quad \cdot {}_5\phi_4\left[\begin{matrix} q^{-j}, q^{j+2b}, q^b, q^b e^{i\theta}, q^b e^{-i\theta} \\ q^{2b}, q^{b+1/2}, -q^{b+1/2}, -q^b \end{matrix}; q, q\right].
\end{aligned} \tag{3.13}
$$

By Andrews' [1, Thm. 1] q-analogue of Watson's ${}_3F_2$ summation formula

$$
{}_4\phi_3\left[\begin{matrix} a, b, c^{1/2}, -c^{1/2} \\ c, (abq)^{1/2}, -(abq)^{1/2} \end{matrix}; q, q\right] = \frac{a^{n/2}(aq, bq, cq/a, cq/b; q^2)_\infty}{(q, abq, cq, cq/ab; q^2)_\infty}, \tag{3.14}
$$

where $b = q^{-n}$ and n is a nonnegative integer, the ${}_4\phi_3$ series in (3.13) equals zero when $n - j$ is odd and equals

$$\frac{(q, q^{a-2b}; q^2)_k}{(q^{2n-4k+a+1}, q^{2n-4k+2b+2}; q^2)_k} q^{2k(n-2k+b+1/2)}$$

when $n - j = 2k$ and $k = 0, 1, \cdots$. Hence, from (3.13) and (3.8),

$$
\begin{aligned}
&{}_5\phi_4\left[\begin{matrix} q^{-n}, q^{n+a}, q^b, q^b e^{i\theta}, q^b e^{-i\theta} \\ q^{2b}, q^{(a+1)/2}, -q^{(a+1)/2}, -q^b \end{matrix}; q, q\right] \\
&\quad = \sum_{k=0}^{[n/2]} \frac{(-1)^n (q^{-n}, q^{n+a}, q^{b+1/2}, -q^{b+1/2}; q)_{n-2k}}{(q, q^{(a+1)/2}, -q^{(a+1)/2}, q^{n-2k+2b}; q)_{n-2k}} \\
&\quad \cdot \frac{(q, q^{a-2b}; q^2)_k}{(q^{2n-4k+a+1}, q^{2n-4k+2b+2}; q^2)_k} q^{2k(n-2k+b+1/2)+(n-2k)(n-2k+1)/2} \\
&\quad \cdot \left\{ {}_4\phi_3\left[\begin{matrix} q^{2k-n}, q^{n-2k+2b}, q^{b/2} e^{i\theta/2}, q^{b/2} e^{-i\theta/2} \\ q^{b+1/2}, -q^{b+1/2}, -q^b \end{matrix}; q, q\right] \right\}^2.
\end{aligned} \tag{3.15}
$$

Since $(-1)^n(q^{-n};q)_{n-2k} \geq 0$, it is clear from (3.15) that

$$(3.16)\qquad {}_5\phi_4\left[\begin{matrix} q^{-n}, q^{n+a}, q^b, q^b e^{i\theta}, q^b e^{-i\theta} \\ q^{2b}, q^{(a+1)/2}, -q^{(a+1)/2}, -q^b \end{matrix}; q, q\right] \geq 0$$

when $a \geq 2b > -1$ and $0 < q < 1$, which gives a q-extension of (3.6) and hence of the inequalities (1.2) used by de Branges in his proof of the Bieberbach, Robertson, and Milin conjectures.

Another q-extension of (3.6) can be derived by observing that from (3.12) we have

$$(3.17)\quad {}_6\phi_5\left[\begin{matrix} q^{-n}, q^{n+a}, q^b, -q^b, q^{a/2}e^{i\theta}, q^{1/2\,a}e^{-i\theta} \\ q^{2b}, q^{(a+1)/2}, -q^{(a+1)/2}, -q^{a/2}, -q^{a/2} \end{matrix}; q, q\right]$$

$$= \sum_{j=0}^{n} \frac{(q^{-n}, q^{n+a}, q^{a/2}, q^{a/2}e^{i\theta}, q^{a/2}e^{-i\theta}; q)_j}{(q, q^{(a+1)/2}, -q^{(a+1)/2}, -q^{a/2}, q^{j+a-1}; q)_j}(-1)^j q^{j+\binom{j}{2}}$$

$$\cdot {}_5\phi_4\left[\begin{matrix} q^{j-n}, q^{n+j+a}, q^{j+a/2}, q^{j+a/2}e^{i\theta}, q^{j+a/2}e^{-i\theta} \\ q^{2j+a}, q^{j+(a+1)/2}, -q^{j+(a+1)/2}, -q^{j+a/2} \end{matrix}; q, q\right]$$

$$\cdot {}_4\phi_3\left[\begin{matrix} q^{-j}, q^{j+a-1}, q^b, -q^b \\ q^{2b}, q^{a/2}, -q^{a/2} \end{matrix}; q, q\right]$$

$$= \sum_{k=0}^{[n/2]} \frac{(q^{-n}, q^{n+a}, q^{a/2}, q^{a/2}e^{i\theta}, q^{a/2}e^{-i\theta}; q)_{2k}}{(q, q^{(a+1)/2}, -q^{(a+1)/2}, -q^{\frac{1}{2}a}, q^{2k+a-1}; q)_{2k}}$$

$$\cdot \frac{(q, q^{a-2b}; q^2)_k}{(q^a, q^{2b+1}; q^2)_k} q^{2k^2+k+2kb}$$

$$\cdot \left\{ {}_4\phi_3\left[\begin{matrix} q^{2k-n}, q^{n+2k+a}, q^{k+a/4}e^{i\theta/2}, q^{k+a/4}e^{-i\theta/2} \\ q^{2k+(a+1)/2}, -q^{2k+(a+1)/2}, -q^{2k+a/2} \end{matrix}; q, q\right] \right\}^2$$

by (3.14) and (3.8). Hence,

$$(3.18)\qquad {}_6\phi_5\left[\begin{matrix} q^{-n}, q^{n+a}, q^b, -q^b, q^{a/2}e^{i\theta}, q^{a/2}e^{-i\theta} \\ q^{2b}, q^{(a+1)/2}, -q^{(a+1)/2}, -q^{a/2}, -q^{a/2} \end{matrix}; q, q\right] \geq 0$$

when $a \geq 2b > -1$ and $0 < q < 1$, which is a q-extension of (3.6) that is different from (3.16). The expansions (8.12) and (8.17) in [18] are special cases of the $q \uparrow 1$ limit cases of (3.15) and (3.17), respectively, when (2.15) and [18, (8.10)] are used.

A q-extension of (3.9) can be derived by using (3.12) and (2.15) to obtain the expansion

(3.19)

$${}_7\phi_6\left[\begin{matrix} q^{-n}, q^{n+\alpha+2}, q^{(\alpha+1)/2}, q^{(\alpha+1)/2}e^{2i\theta}, q^{(\alpha+1)/2}e^{-2i\theta}, q^{(\alpha+2)/2}e^{2i\tau}, q^{(\alpha+2)/2}e^{-2i\tau} \\ q^{\alpha+1}, q^{(\alpha+3)/2}, -q^{(\alpha+3)/2}, -q^{(\alpha+2)/2}, -q^{(\alpha+2)/2}, -q^{(\alpha+1)/2} \end{matrix}; q, q\right]$$

$$= \sum_{j=0}^{n} \frac{(q^{-n}, q^{n+\alpha+2}, q^{(\alpha+2)/2}, q^{(\alpha+2)/2}e^{2i\tau}, q^{(\alpha+2)/2}e^{-2i\tau}; q)_j}{(q, q^{(\alpha+3)/2}, -q^{(\alpha+3)/2}, -q^{(\alpha+2)/2}, q^{j+\alpha+1}; q)_j}(-1)^j q^{j+\binom{j}{2}}$$

$$\cdot \left\{ \frac{(q;q)_j (q;q)_{n-j}}{(q^{\alpha+1};q)_j (q^{2j+\alpha+2};q)_{n-j}} q^{1/2\,(j+\alpha+3/2)} \right.$$

$$\left. \cdot C_j(\cos\theta; q^{(\alpha+1)/2} \mid q) C_{n-j}(\cos\tau; q^{j+(\alpha+2)/2} \mid q) \right\}^2,$$

which is clearly nonnegative for real θ and τ when $\alpha > -2$. The case $\alpha = -2$ can be handled as a limit case of $(q^{\alpha+2};q)_n$ times the ${}_7\phi_6$ series in (3.19); see [4, p. 720] for the hypergeometric case.

Additional nonnegative sums and, in particular, the nonnegativity of q-extensions of the sums of Jacobi polynomials in [18, (8.19), (8.20), (8.22)] will be considered in [23].

4. q-Extensions of de Branges' differential equations. From (3.3) and the identity $(a)_n = \Gamma(n+a)/\Gamma(a)$ it follows that

(4.1)
$$\sigma_n(t) = c\frac{n!(2\nu+2\lambda+1)_{n+r}4^{-n}t^{-n}}{(r-n)!(\nu+1)_n(2\nu+1)_n(\nu+\lambda+1)_n} \cdot {}_4F_3\left[\begin{matrix} n-r, n+r+2\nu+2\lambda+1, n+\nu+1/2, n+2\nu \\ 2n+2\nu+1, n+\nu+\lambda+1, n+2\nu+1 \end{matrix}; t^{-1}\right]$$

for $n = 1, \cdots, r$ with $c = \Gamma(2\nu+2\lambda+1)/[\Gamma(\nu+1)\Gamma(2\nu+1)\Gamma(\nu+\lambda+1)]$. In view of the limit (3.7), we set $t = 2/(1-x)$ and consider the functions

$$\tau_n(x) = c^{-1}\sigma_n\left(\frac{2}{1-x}\right). \tag{4.2}$$

Then de Branges' differential equation (3.4) is equivalent to

$$\frac{n}{n+2\nu}\tau_n(x) + \frac{1-x}{n+2\nu}\tau_n'(x) = \frac{n+2\nu+1}{n+1}\tau_{n+1}(x) - \frac{1-x}{n+1}\tau_{n+1}'(x) \tag{4.3}$$

which can be rewritten in the form

$$\frac{d}{dx}[(1-x)^{-n}\tau_n(x)] = -\frac{n+2\nu}{n+1}(1-x)^{-2n-2\nu-1}\frac{d}{dx}[(1-x)^{n+2\nu+1}\tau_{n+1}(x)]. \tag{4.4}$$

Let $0 < q < 1$, $x = \cos\theta$ and let a, α, λ, ν, θ be real numbers. A q-extension of $(1-x)^\alpha$ can be obtained by extending the definition of the q-shifted factorial to

$$(a;q)_\alpha = \frac{(a;q)_\infty}{(aq^\alpha;q)_\infty} \tag{4.5}$$

and observing that, by the q-binomial theorem [38, (3.2.2.11)],

$$\lim_{q\to 1} 2^{-\alpha}(q^a e^{i\theta}, q^a e^{-i\theta};q)_\alpha = (1-x)^\alpha. \tag{4.6}$$

Hence, if we define

$$u_n(x) = A_{n,r}(q^{\nu+1}e^{i\theta}, q^{\nu+1}e^{-i\theta};q)_n \cdot {}_6\phi_5\left[\begin{matrix} q^{n-r}, q^{n+r+2\nu+2\lambda+1}, q^{n+\nu+1/2}, q^{n+2\nu}, q^{n+\nu+1}e^{i\theta}, q^{n+\nu+1}e^{-i\theta} \\ q^{2n+2\nu+1}, q^{n+\nu+\lambda+1}, q^{n+2\nu+1}, -q^{n+\nu+\lambda+1}, -q^{n+\nu+1/2} \end{matrix}; q, q\right] \tag{4.7}$$

with

$$A_{n,r} = \frac{(q;q)_n(q^{2\nu+2\lambda+1};q)_{n+r}4^{-2n}}{(q;q)_{r-n}(q^{\nu+1}, q^{2\nu+1}, q^{\nu+\lambda+1};q)_n}, \qquad n = 1, \cdots, r, \tag{4.8}$$

and $A_{n,r} = 0$ when $n > r$, then

$$\lim_{q\to 1} u_n(x) = \tau_n(x). \tag{4.9}$$

To obtain a q-extension of differentiation that plays the same role for $(ae^{i\theta}, ae^{-i\theta};q)_n$ as d/dx does for x^n, Askey and Wilson [7, p. 33] defined the operators δ_q and D_q by

$$\delta_q f(e^{i\theta}) = f(q^{1/2}e^{i\theta}) - f(q^{-1/2}e^{i\theta}), \tag{4.10}$$

$$D_q h(x) = \frac{\delta_q h(x)}{\delta_q x}, \tag{4.11}$$

where $x = (e^{i\theta} + e^{-i\theta})/2 = \cos\theta$, and observed that

$$\delta_q(ae^{i\theta}, ae^{-i\theta};q)_n = aq^{-1/2}(1-q^n)(e^{i\theta} - e^{-i\theta})(aq^{1/2}e^{i\theta}, aq^{1/2}e^{-i\theta};q)_{n-1} \tag{4.12}$$

and

$$\frac{\delta_q \prod_{k=0}^{n-1}(1 - 2axq^k + a^2q^{2k})}{\delta_q x} = \frac{-2a(1-q^n)}{1-q}\prod_{k=0}^{n-2}(1 - 2axq^{k+1/2} + a^2q^{2k+1}). \tag{4.13}$$

They noted that when $q \to 1$ formula (4.13) becomes

$$\frac{d}{dx}(1 - 2ax + a^2)^n = -2an(1 - 2ax + a^2)^{n-1} \tag{4.14}$$

and more generally,

$$\lim_{q\to 1} D_q h(x) = \lim_{q\to 1}\frac{\delta_q h(x)}{\delta_q x} = \frac{df(x)}{dx}. \tag{4.15}$$

To derive a q-extension of (4.4) and the inequalities (3.5), first observe that (4.12) extends to

$$\delta_q(ae^{i\theta}, ae^{-i\theta};q)_\alpha = aq^{-1/2}(1-q^\alpha)(e^{i\theta} - e^{-i\theta})(aq^{1/2}e^{i\theta}, aq^{1/2}e^{-i\theta};q)_{\alpha-1}, \tag{4.16}$$

which gives

$$D_q(ae^{i\theta}, ae^{-i\theta};q)_\alpha = \frac{-2a(1-q^\alpha)}{1-q}(aq^{1/2}e^{i\theta}, aq^{1/2}e^{-i\theta};q)_{\alpha-1}. \tag{4.17}$$

Hence, corresponding to the inequality (3.5), we have that

$$\begin{aligned} &D_q[(q^{1-\nu}e^{i\theta}, q^{1-\nu}e^{-i\theta};q)_{2\nu}u_n(x)] \\ &= A_{n,r}\sum_{k=0}^{r-n}\frac{(q^{n-r}, q^{n+r+2\nu+2\lambda+1}, q^{n+\nu+1/2}, q^{n+2\nu};q)_k q^k}{(q, q^{2n+2\nu+1}, q^{n+\nu+\lambda+1}, q^{n+2\nu+1}, -q^{n+\nu+\lambda+1}, -q^{n+\nu+1/2};q)_k} \\ &\quad \cdot D_q[(q^{1-\nu}e^{i\theta}, q^{1-\nu}e^{-i\theta};q)_{n+k+2\nu}] \\ &= -\frac{2(1-q^{n+2\nu})}{1-q}q^{1-\nu}A_{n,r}(q^{3/2-\nu}e^{i\theta}, q^{3/2-\nu}e^{-i\theta};q)_{n+2\nu-1} \\ &\quad \cdot {}_5\phi_4\left[\begin{matrix} q^{n-r}, q^{n+r+2\nu+2\lambda+1}, q^{n+\nu+1/2}, q^{n+\nu+1/2}e^{i\theta}, q^{n+\nu+1/2}e^{-i\theta} \\ q^{2n+2\nu+1}, q^{n+\nu+\lambda+1}, -q^{n+\nu+\lambda+1}, -q^{n+\nu+1/2}\end{matrix};q,q\right] \le 0 \end{aligned} \tag{4.18}$$

by (3.16), when $\nu > -\frac{1}{2}$, $\lambda \ge 0$ and $n = 1, \cdots, r$. This explains why we chose the ${}_6\phi_5$ in (4.7).

To derive a q-extension of (4.4) notice that, corresponding to the left side of (4.4), we have

(4.19)

$$D_q[(q^{n+\nu+1}e^{i\theta}, q^{n+\nu+1}e^{-i\theta}; q)_{-n}u_n(x)] = A_{n,r}$$
$$\cdot D_q\left({}_6\phi_5\left[\begin{matrix} q^{n-r}, q^{n+r+2\nu+2\lambda+1}, q^{n+\nu+1/2}, q^{n+2\nu}, q^{n+\nu+1}e^{i\theta}, q^{n+\nu+1}e^{-i\theta} \\ q^{2n+2\nu+1}, q^{n+\nu+\lambda+1}, q^{n+2\nu+1}, -q^{n+\nu+\lambda+1}, -q^{n+\nu+1/2} \end{matrix}; q, q\right]\right)$$
$$= \frac{(1-q^{n-r})(1-q^{n+r+2\nu+2\lambda+1})(1-q^{n+\nu+1/2})(1-q^{n+2\nu})(-2)q^{n+\nu+2}}{(1-q)(1-q^{2n+2\nu+1})(1-q^{n+\nu+\lambda+1})(1-q^{n+2\nu+1})(1+q^{n+\nu+\lambda+1})(1+q^{n+\nu+1/2})} A_{n,r}$$
$$\cdot {}_6\phi_5\left[\begin{matrix} q^{n+1-r}, q^{n+r+2\nu+2\lambda+2}, q^{n+\nu+3/2}, q^{n+2\nu+1}, q^{n+\nu+3/2}e^{i\theta}, q^{n+\nu+3/2}e^{-i\theta} \\ q^{2n+2\nu+2}, q^{n+\nu+\lambda+2}, q^{n+2\nu+2}, -q^{n+\nu+\lambda+2}, -q^{n+\nu+3/2} \end{matrix}; q, q\right].$$

Similarly,

(4.20)

$$(q^{n+\nu+3/2}e^{i\theta}, q^{n+\nu+3/2}e^{-i\theta}; q)_{-2n-2\nu-1} D_q[(q^{-n-\nu}e^{i\theta}, q^{-n-\nu}e^{-i\theta}; q)_{n+2\nu+1}u_{n+1}(x)]$$
$$= -\frac{2(1-q^{2n+2\nu+2})}{1-q} q^{-n-\nu} A_{n+1,r}$$
$$\cdot {}_6\phi_5\left[\begin{matrix} q^{n+1-r}, q^{n+r+2\nu+2\lambda+2}, q^{n+\nu+3/2}, q^{n+2\nu+1}, q^{n+\nu+3/2}e^{i\theta}, q^{n+\nu+3/2}e^{-i\theta} \\ q^{2n+2\nu+2}, q^{n+\nu+\lambda+2}, q^{n+2\nu+2}, -q^{n+\nu+\lambda+2}, -q^{n+\nu+3/2} \end{matrix}; q, q\right],$$

which, combined with (4.19), gives the following q-extension of (4.4)

(4.21) $$D_q[(q^{n+\nu+1}e^{i\theta}, q^{n+\nu+1}e^{-i\theta}; q)_{-n}u_n(x)]$$
$$= -\frac{1-q^{n+2\nu}}{1-q^{n+1}} B_{n,r}(q^{n+\nu+3/2}e^{i\theta}, q^{n+\nu+3/2}e^{-i\theta}; q)_{-2n-2\nu-1}$$
$$\cdot D_q[(q^{-n-\nu}e^{i\theta}, q^{-n-\nu}e^{-i\theta}; q)_{n+2\nu+1}u_{n+1}(x)]$$

with

(4.22) $$B_{n,r} = \frac{16q^{3n-r+2\nu+2}}{(1+q^{n+\nu+\lambda+1})(1+q^{n+\nu+1})(1+q^{n+\nu+1/2})^2}.$$

Clearly, $B_{n,r} \to 1$ and (4.21) tends to (4.4) as $q \to 1$; but, unlike (4.4), the difference equation (4.21) depends on r. However, if we consider following positive multiple of u_n

(4.23) $$U_n(x) = \frac{4^{2n}q^{3n^2/2+n(2\nu+1/2-r)}}{(-q^{\nu+\lambda+1}, -q^{\nu+1}, -q^{\nu+1/2}, -q^{\nu+1/2}; q)_n} u_n(x),$$

we find that it satisfies the difference equation

(4.24) $$D_q[(q^{n+\nu+1}e^{i\theta}, q^{n+\nu+1}e^{-i\theta}; q)_{-n}U_n(x)]$$
$$= -\frac{1-q^{n+2\nu}}{1-q^{n+1}} (q^{n+\nu+3/2}e^{i\theta}, q^{n+\nu+3/2}e^{-i\theta}; q)_{-2n-2\nu-1}$$
$$\cdot D_q[(q^{-n-\nu}e^{i\theta}, q^{-n-\nu}e^{-i\theta}; q)_{n+2\nu+1}U_{n+1}(x)],$$

which is independent of r.

Similarly, setting

(4.25)

$$V_n(x) = C_{n,r}(q^{\nu+\lambda+1}e^{i\theta}, q^{\nu+\lambda+1}e^{-i\theta}; q)_n$$
$$\cdot {}_7\phi_6\left[\begin{matrix} q^{n-r}, q^{n+r+2\nu+2\lambda+1}, q^{n+\nu+1/2}, -q^{n+\nu+1/2}, q^{n+2\nu}, q^{n+\nu+\lambda+1}e^{i\theta}, q^{n+\nu+\lambda+1}e^{-i\theta} \\ q^{2n+2\nu+1}, q^{n+\nu+\lambda+1}, -q^{n+\nu+\lambda+1}, q^{n+2\nu+1}, -q^{n+\nu+\lambda+1/2}, -q^{n+\nu+\lambda+1/2} \end{matrix}; q, q\right]$$

with

(4.26)

$$C_{n,r} = \frac{(q;q)_n(q^{2\nu+2\lambda+1};q)_{n+r}q^{3n^2/2+n(2\nu+1/2-r)}}{(q;q)_{r-n}(q^{\nu+1}, q^{2\nu+1}, q^{\nu+\lambda+1}, -q^{\nu+1}, -q^{\nu+\lambda+1}, -q^{\nu+\lambda+1/2}, -q^{\nu+\lambda+1/2};q)_n}$$

when $n = 1, \cdots, r$ and $C_{n,r} = 0$ when $n > r$, we obtain that $V_n(x) \to \tau_n(x)$ as $q \to 1$, $V_n(x)$ satisfies the following q-extension of (4.4)

(4.27)
$$D_q[(q^{n+\nu+\lambda+1}e^{i\theta}, q^{n+\nu+\lambda+1}e^{-i\theta}; q)_{-n}V_n(x)]$$
$$= -\frac{1-q^{n+2\nu}}{1-q^{n+1}}(q^{n+\nu+\lambda+3/2}e^{i\theta}, q^{n+\nu+\lambda+3/2}e^{-i\theta}; q)_{-2n-2\nu-1}$$
$$\cdot D_q[(q^{-n-\nu+\lambda}e^{i\theta}, q^{-n-\nu+\lambda}e^{-i\theta}; q)_{n+2\nu+1}V_{n+1}(x)]$$

and, by (3.18),

(4.28)

$$D_q[(q^{1+\lambda-\nu}e^{i\theta}, q^{1+\lambda-\nu}e^{-i\theta}; q)_{2\nu}V_n(x)] = -\frac{2(1-q^{n+2\nu})}{1-q}q^{1+\lambda-\nu}C_{n,r}$$
$$\cdot (q^{\lambda+3/2-\nu}e^{i\theta}, q^{\lambda+3/2-\nu}e^{-i\theta}; q)_{n+2\nu-1}$$
$$\cdot {}_6\phi_5\left[\begin{matrix} q^{n-r}, q^{n+r+2\nu+2\lambda+1}, q^{n+\nu+1/2}, -q^{n+\nu+1/2}, q^{n+\nu+\lambda+1/2}e^{i\theta}, q^{n+\nu+\lambda+1/2}e^{-i\theta} \\ q^{2n+2\nu+1}, q^{n+\nu+\lambda+1}, -q^{n+\nu+\lambda+1}, -q^{n+\nu+\lambda+1/2}, -q^{n+\nu+\lambda+1/2} \end{matrix}; q, q\right] \le 0$$

when $\nu > -1/2$, $\lambda \ge 0$ and $n = 1, \cdots, r$.

The q-extensions of de Branges' inequalities and differential equations contained in this paper suggest that it might be possible to extend some of the other parts of his proof of the Bieberbach, Robertson, and Milin conjectures. Besides (1.2) and (1.7), de Branges also used the fact that if $F(t,z)$ is a Löwner family of Riemann mapping functions, then

$$t\frac{\partial}{\partial t}F(t,z) = \varphi(t,z)z\frac{\partial}{\partial z}F(t,z), \tag{4.29}$$

where $\varphi(t,z)$ is a power series with constant coefficient equal to 1,which represents a function with positive real part in the unit disk for every index t, and the coefficients of $\varphi(t,z)$ are measurable functions of t. q-Extensions of the Löwner [29] theory and of the coefficient estimates for Riemann mapping functions in [12] and [13, Thms. 1–4] are still open. In view of the definition of D_q in (4.11), a prospect for a q-extension of (4.29) is the equation

$$(1-x)D_qG(x,z) = \Phi(x,z)z\frac{\partial}{\partial z}G(x,z) \tag{4.30}$$

or this equation with the partial derivative replaced by a difference operator. For an extremal function that is a q-extension of the Koebe function, the ${}_1F_0$ series representation for the Koebe function

$$k(z) = \frac{z}{(1-z)^2} = z\,{}_1F_0\left[\begin{matrix} 2 \\ - \end{matrix}; z\right] \tag{4.31}$$

suggests that a natural choice is the "q-Koebe" function

$$k_q(z) = z\ {}_1\phi_0\left[\begin{matrix} q^2 \\ - \end{matrix}; q, z\right] = \frac{z}{(1-z)(1-qz)}. \tag{4.32}$$

Note that, just as the Koebe function is starlike, $k_q(z)$ is a starlike function when $-1 < q < 1$, which can be shown by using [16, Thm. 2.10] and the positivity of the Poisson kernel for Fourier series.

REFERENCES

[1] G. E. ANDREWS, *On q-analogues of the Watson and Whipple summations*, SIAM J. Math. Anal., 7 (1976), pp. 332–336.

[2] G. E. ANDREWS AND R. ASKEY, *Enumeration of partitions: The role of Eulerian series and q-orthogonal polynomials*, in Higher Combinatorics, M. Aigner and D. Reidel, eds., Dordrecht, the Netherlands, 1977, pp. 3–26.

[3] R. ASKEY, *Orthogonal Polynomials and Special Functions*, Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1975.

[4] R. ASKEY AND G. GASPER, *Positive Jacobi polynomial sums II*, Amer. J. Math., 98 (1976), pp. 709–737.

[5] ———, *Inequalities for polynomials*, in The Bieberbach Conjecture: Proc. of the Symposium on the Occasion of the Proof, A. Baernstein, D. Drasin, P. Duren, and A. Marden, eds., Math. Surveys Monographs 21, American Mathematical Society, Providence, R.I. 1986, pp. 7–32.

[6] R. ASKEY AND M. ISMAIL, *A generalization of ultraspherical polynomials*, in Studies in Pure Mathematics, P. Erdös, ed., Birkhäuser, Basel, 1983, pp. 55–78.

[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 Stechert–Hafner, New York, 1964).

[9] L. BIEBERBACH, *Über die Koeffizienten derjenigen Potenzreihen, welche eine schlichte Abbildung des Einheitskreises vermitteln*, Sitzungsberichte Preuss. Akad. Wiss., Berlin, 1916, pp. 940–955.

[10] L. DE BRANGES, *Quantum Cesàro operators*, in Topics in Functional Analysis, Advances in Mathematics Supplementary Studies, Vol. 3, I. Gohberg and M. Kac, eds., Academic Press, New York, 1978.

[11] ———, *A proof of the Bieberbach conjecture*, USSR Academy of Sciences, Steklov Math. Institute, LOMI, preprint E-5-84, Leningrad, 1984.

[12] ———, *A proof of the Bieberbach conjecture*, Acta Math., 154 (1985), pp. 137–152.

[13] ———, *Powers of Riemann mapping functions*, in The Bieberbach Conjecture: Proc. of the Symposium on the Occasion of the Proof, A. Baernstein, D. Drasin, P. Duren, and A. Marden, eds., Math. Surveys and Monographs, 21, 1986, American Mathematical Society, Providence, R.I., pp. 51–67.

[14] D. M. BRESSOUD, *Linearization and related formulas for q-ultraspherical polynomials*, SIAM J. Math. Anal., 12 (1981), pp. 161–168.

[15] T. CLAUSEN, *Ueber die Fälle, wenn die Reihe von der Form... ein Quadrat von der Form... hat*, J. Reine Angew. Math., 3 (1828), pp. 89–91.

[16] P. L. DUREN, *Univalent Functions*, Springer-Verlag, Berlin, New York, 1983.

[17] J. L. FIELDS AND J. WIMP, *Expansions of hypergeometric functions in hypergometric functions*, Math. Comp., 15 (1961), pp. 390–395.

[18] G. GASPER, *Positivity and special functions*, in Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 375–433.

[19] ———, *Positive sums of the classical orthogonal polynomials*, SIAM J. Math. Anal., 8 (1977), pp. 423–447.

[20] ———, *Rogers' linearization formula for the continuous q-ultraspherical polynomials and quadratic transformation formulas*, SIAM J. Math. Anal., 16 (1985), pp. 1061–1071.

[21] ———, *A short proof of an inequality used by de Branges in his proof of the Bieberbach, Robertson and Milin conjectures*, Complex Variables, Theory Appl., 7 (1986), pp. 45–50.
[22] ———, *Summation, transformation, and expansion formulas for bibasic series*, Trans. Amer. Math. Soc., to appear.
[23] ———, *Expansion formulas for basic hypergeometric series and nonnegative sums of Askey-Wilson polynomials*, to appear.
[24] G. GASPER AND M. RAHMAN, *Product formulas of Watson, Bailey and Bateman types and positivity of the Poisson kernel for q-Racah polynomials*, SIAM J. Math. Anal., 15 (1984), pp. 768–789.
[25] ———, *Basic Hypergeometric Series*, Cambridge University Press, to appear.
[26] ———, *A nonterminating q-Clausen formula and some related product formulas*, SIAM J. Math. Anal., 20 (1989), to appear.
[27] F. H. JACKSON, *The q^θ equations whose solutions are products of solutions of q^θ equations of lower order*, Quart. J. Math. Oxford, 11 (1940), pp. 1–17.
[28] ———, *Certain q-identities*, Quart. J. Math. Oxford, 12 (1941), pp. 167–172.
[29] K. LÖWNER, *Untersuchungen über schlichte konforme Abbildungen des Einheitskreises. I*, Math. Ann., 89 (1923), pp. 103–121.
[30] I. M. MILIN, *Univalent Functions and Orthogonal Systems*, Transl. Math. Monographs 49, American Mathematical Society, Providence, R.I., 1977.
[31] M. RAHMAN, *The linearization of the product of continuous q-Jacobi polynomials*, Canad. J. Math., 23 (1981), pp. 961–987.
[32] M. RAHMAN AND A. VERMA, *Product and addition formulas for the continuous q-ultraspherical polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 1461–1474.
[33] M. S. ROBERTSON, *A remark on the odd schlicht functions*, Bull. Amer. Math. Soc., 42 (1936), pp. 366–370.
[34] L. J. ROGERS, *On the expansion of some infinite products*, Proc. London Math. Soc., 24 (1893), pp. 337–352.
[35] ———, *Second memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 25 (1894), pp. 318–343.
[36] ———, *Third memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 26 (1895), pp. 15–32.
[37] D. B. SEARS, *On the transformation theory of basic hypergeometric functions*, Proc. London Math. Soc. (2) 53 (1951), pp. 158–180.
[38] L. J. SLATER, *Generalized Hypergometric Functions*, Cambridge University Press, London, 1966.
[39] A. VERMA, *Certain expansions of the basic hypergeometric functions*, Math. Comp., 20 (1966), pp. 151–157.
[40] ———, *Some transformations of series with arbitrary terms*, Instituto Lombardo Rend. Sc. A 106 (1972), pp. 342–353.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1035–1054, September 1989

THE EXISTENCE AND BEHAVIOR OF VISCOUS STRUCTURE FOR PLANE DETONATION WAVES*

DAVID H. WAGNER†

Abstract. A necessary condition and a sufficient condition are proved for the existence of steady plane wave solutions to the Navier–Stokes equations for a reacting gas. These solutions represent plane detonation waves, and converge to ZND detonation waves as the viscosity, heat conductivity, and species diffusion rates tend to zero. It is assumed that the Prandtl number is $\frac{3}{4}$, but arbitrary Lewis numbers are permitted. No assumption is made concerning the activation energy.

It is shown that the stagnation enthalpy and the entropy flux are always monotone for such solutions, and that the mass density and pressure are nearly always not monotone, as predicted by the ZND theory.

In certain parameter ranges, typically that of large diffusion, many of these waves have the appearance of a weak detonation followed by an inert shock wave. This confirms a phenomenon observed in numerical calculations and in a model system by Colella, Majda, and Roytburd.

Key words. shock waves, detonation waves, combustion

AMS(MOS) subject classifications. 76L05, 80A32, 80A25, 35L65

1. Introduction.

Detonation waves are compressive, exothermically reacting shock waves. One of the curiosities of combustion theory is that there also exist expansive "shock waves" known as *deflagration waves*, which will not be discussed in this paper. We will give a mathematically rigorous, but simple, discussion of the *viscous structure* of plane detonation waves.

We begin with a brief discussion of the inviscid theory, known as the Chapman–Jouguet (CJ) theory. By means of a Galilean transformation we may reduce the problem to one of studying steady plane waves. If we assume that the thickness of the reaction zone is zero, if we neglect all diffusion effects such as viscosity, heat conduction, and diffusion of species, and any external forces such as gravity, and if we look for steady plane waves, then we obtain the following system of differential equations:

$$
\begin{aligned}
&\text{(a)} \quad (\rho u)_x = 0,\\
&\text{(b)} \quad \left[\rho u^2 + p(\rho, T)\right]_x = 0,\\
&\text{(c)} \quad \left((\rho(u^2/2 + e(\rho, T, Y)) + p(\rho, T))u\right)_x = 0,\\
&\text{(d)} \quad (\rho u Y)_x = -\rho u(Y_-)\delta(x - x_0).
\end{aligned}
\tag{1.1}
$$

Here x is a space coordinate in the direction normal to the wave, x_0 is the location of the wave, and ρ, T, u, p, e, and Y are the mass density, temperature, x component of velocity, pressure, specific internal energy, and mass fraction of the reactant, respectively. Y_- is the unburned value of Y. As is standard practice, we have represented the extremely complicated chemical reaction by a simplified, one-step chemistry: reactant $\rightarrow$ product. From (1.1a) we see that the mass flux, ρu, has a constant value; we denote this value by m. The fluxes of momentum (1.1b) and energy (1.1c) are also constant; from this fact we obtain the *Rankine–Hugoniot conditions* for a shock wave, which

*Received by the editors January 29, 1988; accepted for publication January 17, 1989.

†Department of Mathematics, University of Houston, Houston, Texas 77204-3476. This research was supported by National Science Foundation grant DMS-8601917 and Air Force Office of Scientific Research grant AFOSR 86-0218.

in the inviscid theory is represented by a jump discontinuity in the unknowns. The difference between inert gas dynamics, and the exothermic reactive theory discussed here, lies in the fact that Y varies from a positive value on the unburned side of the wave, which we take to lie on the left side, to a zero value on the burned, or right side. Because the internal energy e depends on Y, the change in Y causes the classical Hugoniot curve (the solution locus of (1.1c)) of gas dynamics to move. As a consequence, we find that, for a given value of m, a given shock state on the left may now be connected by a shock wave to two possible states on the right, except for certain critical values of m for which there is a unique burned state — the *Chapman–Jouguet point* — see Fig. 1. In addition, the curve of possible burned states, parameterized by m, has two components. One component, corresponding to compressive waves, is called the detonation branch, and the other component, corresponding to expansive waves, is called the deflagration branch. By way of contrast, in an inert gas, for a given value of m, a state is usually connected to only one state on the right, and the curve of possible terminal shock states is usually connected.

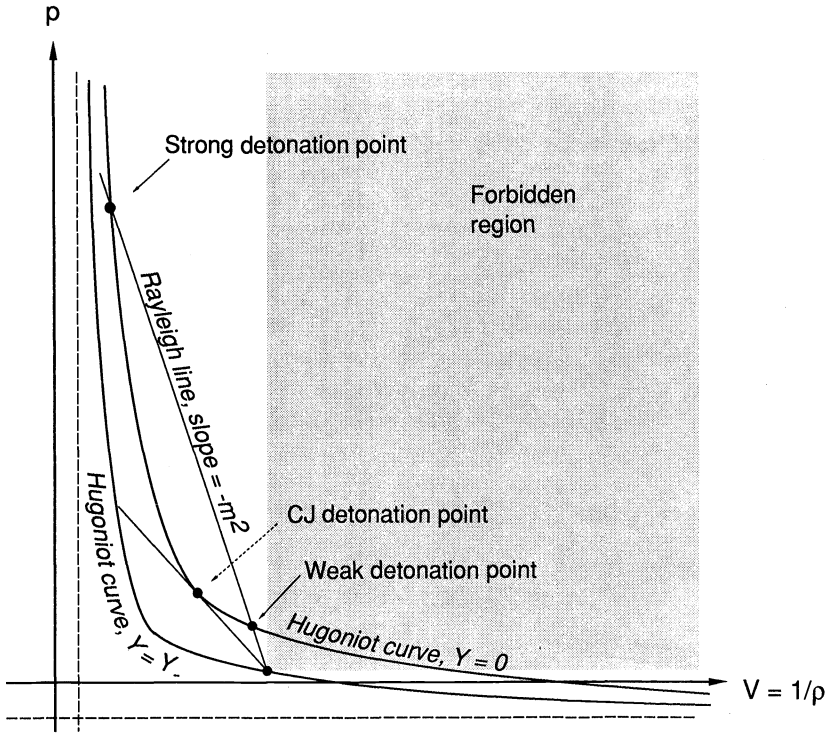


FIG. 1. *The Chapman–Jouguet diagram.*

The combustive shock waves of the CJ theory are classified as follows. A wave connecting the unburned state to the closer detonation point is called a *weak* detonation wave, and a connection to the farther detonation point is called a *strong* detonation. A detonation wave terminating at the Chapman–Jouguet point is called a *Chapman–Jouguet* detonation. Deflagration waves are similarly classified. For the exothermic, irreversible reactions considered here, strong deflagrations violate the second law of thermodynamics and are unphysical, and weak detonations are rare. If we permit an endothermic region then strong deflagrations and weak detonations are possible and

perhaps even probable [FD].

The CJ theory for detonation waves is useful for deriving the Rankine–Hugoniot conditions, and for classifying the types of wave. However, this theory is physically flawed, because in reality the reaction zone is much thicker than the shock layer. This is due to the fact that the chemical reaction depends on molecular collisions and requires a distance much longer than the mean free path to achieve significant completion. The shock layer, however, has been experimentally observed to be several mean free paths thick. Consequently the appropriate inviscid model is the one developed independently by Zel'dovich, von Neumann, and Döring [Z], [ZK], [N1], [N2], [D], and which is known as the ZND model. In this model equation (1.1d) is replaced by a similar equation, but with a finite reaction rate:

$$(\rho u Y)_x = -r(\rho, Y, T) \tag{1.1d'}$$

For our purposes it is reasonable to assume that the reaction rate function r is continuous, nonnegative, and monotone in each variable. Our mathematical treatment will require that we assume that r vanishes whenever the temperature T is less than a given *ignition temperature* T_i. For a known reaction rate r (1.1a–c, d′) can be solved explicitly; the only detonation wave solutions are strong or CJ detonations. These waves, which are known as ZND waves, begin with a jump discontinuity which is an inert shock wave. This shock wave heats the gas above the ignition temperature; the reaction proceeds, with the velocity and temperature following a curve of equilibrium states for (1.1b, c), parameterized by Y. One of the interesting features of these waves is the peak in the pressure and density which is known as the *von Neumann spike*. By way of contrast, in inert shock waves the pressure and density are usually monotone [Gi].

The equations of inert, inviscid, non-heat-conducting gas dynamics are an example of a nonlinear hyperbolic system of conservation laws. In the theory for such systems it is standard practice to set admissibility criteria to distinguish physical from unphysical shock waves. One of the criteria in which much faith is put is to accept a shock wave as physical if it is *structurally stable*. A shock wave is structurally stable if it is the limit of solutions to models which include more physical effects, such as viscosity and heat conduction, as these models tend to the original inviscid model in which these effects are neglected. For steady plane detonation waves the effects of viscosity, heat conduction, and species diffusion may be considered to obtain the (steady) *reacting compressible Navier–Stokes equations*:

$$\begin{aligned}
&\text{(a)} \quad (\rho u)_x = 0, \\
&\text{(b)} \quad \left(\rho u^2 + p(\rho, T)\right)_x = (\mu u_x)_x, \\
&\text{(c)} \quad \left[\left(\rho\left(u^2/2 + e(\rho, T, Y)\right) + p(\rho, T)\right) u\right]_x = (\lambda T_x)_x + (\mu u u_x)_x + (q\rho D Y_x)_x, \\
&\text{(d)} \quad (\rho u Y)_x = (\rho D Y_x)_x - r(\rho, T, Y).
\end{aligned} \tag{1.2}$$

Here μ is the coefficient of viscosity, λ is the heat conductivity, D is the diffusion rate for the reactant, and q is the difference in the heats of formation of the reactant and the product [Wi].

In this paper we prove a necessary condition and a sufficient condition for the existence of heteroclinic solutions of (1.2) which extend from an unburned state at $x = -\infty$ to the strong detonation point at $x = \infty$. These conditions also apply to the Chapman–Jouguet detonation. See (4.8) and (5.3). For simplicity we restrict our

attention to the case where the *Prandtl number* is $\frac{3}{4}$ ($\mu = \lambda/c_p$). In the limit as λ, μ, and D tend to zero (with other parameters fixed) these solutions tend to the ZND wave. Thus the ZND wave is structurally stable to this particular perturbation of the model.

For all of these solutions the *stagnation enthalpy* $H = c_pT + u^2/2$ is monotone, as is the *entropy flux*: $mS - \lambda T_x/T$. For most of the strong detonation waves the density and the pressure attain their maxima in the interior of the wave; this corresponds to the von Neumann spike which occurs in the ZND wave. However, for a certain parameter range, namely whenever

$$\frac{L}{2}\left(1 - \sqrt{\frac{1 + 4Dk\phi(T^*)}{u^{*2}}}\right) > \frac{1}{\gamma}\left(1 - M^{*-2}\right),$$

where $L = \lambda/\rho Dc_p$ is the *Lewis number*, M^* and T^* are the Mach number and temperature at the strong detonation point, and $k\phi(T^*)$ is the reaction rate as defined in (2.9), there exists a continuum of solutions which look like a weak detonation followed by a gas dynamic shock wave. For these waves the pressure and temperature are monotone. This pathological behavior has been noted before in [Wo2], [FD], [LL], [HoSt], and in numerical computations of solutions of the time-dependent Navier–Stokes equations for a reacting gas [CMR]. In these numerical computations it was observed that the weak detonation-shock wave solutions are dynamically stable as solutions of the time-dependent equations in one space dimension.

Since Zel'dovich, von Neumann, and Döring [Z], [N1], [N2], [D] described the typical plane detonation heuristically as an inert shock wave followed by a deflagration, there have been a number of papers on the structure problem for plane detonations. A common assumption has been that the Prandtl number is $\frac{3}{4}$ and that the Lewis number is 1. Under these assumptions Hirschfelder and Curtiss gave a good analysis of the behavior of structure profiles [HC], and Wood proved the existence of structure for "small" reaction rates [Wo1]. The approach taken here has much in common with Wood's, except that we have been more precise in our analysis, our results are stronger and more general, and we give explicit and fairly sharp statements of just how small the rate parameter must be.

A typical expression for the reaction rate is given by the Arrhenius law:

$$r(\rho, T, Y) = k\rho \cdot Ye^{-\theta/T}, \tag{1.3}$$

where θ is the activation energy E/R. In mathematical combustion theory it is standard practice to simplify models such as (1.2) by taking an appropriate distinguished limit as k and θ tend to infinity. This has the effect of reducing the thickness of the reaction zone to zero. Bush and Fendell [BF] gave a description of CJ detonations using asymptotic expansions in the limit of infinite activation energy. Stewart and Holmes proved the existence of viscous structure for (1.2), assuming a large, finite activation energy [HoSt]. Lu and Ludford gave a simplified analysis of weak, strong, and CJ detonations in the infinite activation energy limit [LL].

It is desirable to understand the existence of structure for plane detonation waves with no assumptions concerning activation energy, independent of any desire for mathematical generality. Because (under the ZND hypothesis) the reaction zone is much thicker than the shock layer, this problem should be studied for activation energies that are small compared to the reciprocals of the heat conductivity, viscosity, and species diffusion rate.

Gardner proved the existence of travelling plane detonation wave solutions to the Lagrangian reacting compressible Navier–Stokes equations [Ga]. He made no assumption on the Lewis and Prandtl numbers. However, he omitted the species diffusion term from the energy balance equation, and this term is not usually neglected. It may be that the effect of this term on the solution is very small. However, we will demonstrate in this paper that inclusion of this term permits a much more natural treatment of the problem and better bounds on the solution.

The remainder of the paper is organized as follows. In §2 we explain our assumptions in more detail, and we show that under these assumptions, and the assumption that the Lewis number is 1, that (1.2) is reduced to a system of three differential equations. This material has been extracted and specialized from [Wi]. It is included for clarity and completeness of exposition. In §3 we present the simple topological argument that proves the existence of viscous structure for steady plane detonation waves. In §4 we prove the estimate that makes the topological argument work, and we conclude this section with a sufficient condition for the existence of structure. In §5 we prove, using simple energy estimates on the stagnation enthalpy, a necessary condition for the existence of structure; if this condition is not satisfied then viscous structure does not exist. In §6 we give the generalization to arbitrary Lewis numbers. In §7 we discuss the behavior of the solutions for various parameter values. In §8 we present a rigorous discussion of the ZND limit. We conclude, in §9, with a proof that the entropy flux is monotone.

2. Reduction to three equations. We make the basic thermodynamic assumption that both the reactant and product satisfy the same ideal gas law, and differ only in their heats of formation. Thus the pressure is independent of Y:

$$p = R\rho T. \tag{2.1}$$

We further assume that the internal energy depends linearly on Y, so that we have:

$$e = c_v T + qY, \tag{2.2}$$

where c_v is the specific heat at constant volume. These assumptions are probably not essential to the results that follow, and they could probably be replaced with more qualitative conditions similar to Weyl's conditions for the equation of state for an inert gas. However, these assumptions are essential to the simplifications that follow.

Observe that (1.2) can be integrated once to yield $\rho u = m = constant$. If for any unknown U we let $U_\pm$ be the limit of U as x tends to $\pm\infty$, we have:

$$\begin{aligned}
&\text{(a)} \quad \mu u_x = m(u - u_\pm) + p - p_\pm \\
&\text{(b)} \quad \lambda T_x + \mu u u_x + q\rho D Y_x \\
&\qquad\qquad = m\left[c_v\,(T - T_\pm) + q\,(Y - Y_\pm) + \frac{1}{2}\left(u^2 - u_\pm^2\right) + R\,(T - T_\pm)\right] \\
&\text{(c)} \quad (\rho D Y_x)_x = mY_x + r(\rho, Y, T)
\end{aligned} \tag{2.3}$$

Let $H = c_p T + u^2/2$ be the stagnation enthalpy. Here $c_p = c_v + R$ and is the specific heat at constant pressure. Then (2.3b) may be rewritten as:

$$(\lambda/c_p)H_x + (\mu - \lambda/c_p)\left(\frac{u^2}{2}\right)_x = m\,(H - H_\pm + q(\epsilon - \epsilon_\pm))\,, \tag{2.4}$$

where $\epsilon = Y - \rho D Y_x/m$ is a *reaction progress variable.* When the Prandtl number is $\frac{3}{4}$, then $\mu = \lambda/c_p$ and we have:

$$\left(\frac{\lambda}{mc_p}\right) H_x = H - H_\pm + q(\epsilon - \epsilon_\pm). \tag{2.5}$$

It is convenient to let y satisfy

$$\frac{dy}{dx} = \frac{mc_p}{\lambda}, \tag{2.6}$$

so that the left-hand side of (2.5) becomes H_y. The system (2.3) reduces to a system of four first-order equations:

$$\begin{aligned}
&\text{(a)} \quad mu_y = m(u - u_\pm) + mR\left(\frac{T}{u} - \frac{T_\pm}{u_\pm}\right)\\
&\text{(b)} \quad H_y = H - H_\pm + q(\epsilon - \epsilon_\pm)\\
&\text{(c)} \quad \epsilon_y = -\frac{\lambda}{m^2 c_p} r(\rho, Y, T)\\
&\text{(d)} \quad Y_y = \frac{\lambda}{\rho D c_p}(Y - \epsilon).
\end{aligned} \tag{2.7}$$

Note that

$$(H - H_+)_y + \frac{\rho D c_p}{\lambda} q Y_y = H - H_+ + qY. \tag{2.8}$$

Thus if the *Lewis number*, $L = \lambda/\rho D c_p$, is 1, we see that the quantity $H - H_+ + q(Y - Y_+)$ satisfies the differential equation $f' = f$ and can be bounded only if it is identically zero. As we are interested only in bounded solutions, we may, in this case, restrict our attention to the plane $H - H_+ + q(Y - Y_+) = 0$, and (2.7) reduces to a system of three equations: (2.7a–c) with Y replaced by $(H_+ - H)/q$. We will not actually need to assume that the Lewis number is 1, however, this case is easier to understand.

Some of the results concerning this system are easier to interpret if we replace u by the *specific volume* $V = 1/\rho = u/m$. In this case (2.7a) becomes:

$$V_y = V - V_\pm + \frac{R}{m^2}\left(\frac{T}{V} - \frac{T_\pm}{V_\pm}\right). \tag{2.7a′}$$

With the exception of our discussion of entropy, we will restrict our attention to the system (2.7a′, b–d) = (2.7) for the remainder of this paper.

Following standard practice, we will resolve the *cold boundary difficulty* by means of an *ignition temperature* assumption. The cold boundary difficulty consists of the fact that the Arrhenius reaction rate (1.3) does not vanish, but is merely very small, at the unburned state. Thus no solution of (2.7) can tend to the unburned state as $x \to \infty$. Clearly this problem stems more from our unphysical, infinite domain than from any flaw in the reaction rate. However, as is customary, we will resolve this problem by modifying r so that it is zero for $T < T_i$, where T_i is the *ignition temperature*, and is chosen to be greater than the unburned temperature T_-.

Accordingly we will consider reaction rates of the form:

$$r(\rho, Y, T) = k\rho Y \phi(T), \tag{2.9}$$

where ϕ is a nonnegative Lipschitz monotone function of T, which vanishes for $T < T_i$.

In the next three sections we will usually assume that the Lewis number is 1, for ease of understanding. In §6 we will explain how to generalize to arbitrary Lewis numbers.

3. A picture is worth .... We now describe a region in (V, H, ϵ) space. The topological properties of the flow for (2.7) in this region will imply the existence of the desired structure profiles and will yield other properties as well.

For any bounded solution of (2.7) on which some quantity is strictly monotone, (V, H, ϵ) must tend to a rest state of (2.7) as y tends to $\pm\infty$. From (2.7c) we see that $r(\rho, Y, T)$ must vanish at any rest state. For our modified kinetics this can only happen if $Y = 0$ or $T < T_i$. Rest states satisfying $Y = 0$ are possible *burned states*; those satisfying $T < T_i$ are possible *unburned*, or *fresh gas* states. By (2.7d), ϵ is decreasing, so that the unburned state is at $y = -\infty$ and the burned state is at $y = +\infty$. Thus $Y_+ = 0$ and $\epsilon_+ = (Y - \rho D Y_x/m)_+ = 0$. Note that our choice of the coordinate y, and of V, makes the system (2.7) independent of the sign of $m = \rho u$.

At the unburned state, Y has a given value, Y_-, which, since Y is a mass fraction, is naturally bounded by 1. Since $\epsilon_- = Y_-$, we have, by (2.7b), $H_+ - H_- = qY_-$, which is the total heat per unit mass released by the reaction. Then (2.7a′) yields

$$\begin{aligned} &V_+ - V_- + \frac{R}{m^2}\left(\frac{T_+}{V_+} - \frac{T_-}{V_-}\right) \\ &\quad = (V_+ - V_-)\left(1 - \frac{R}{2c_p}\right) + \frac{R}{m^2 c_p}\left(\frac{H_- + qY_-}{V_+} - \frac{H_-}{V_-}\right) = 0. \end{aligned} \tag{3.1}$$

Let γ be the ratio of specific heats, $c_p/c_v = c_p/(c_p - R)$. Then (3.1) becomes:

$$V_+^2 - \left(V_- + \frac{2(\gamma-1)H_-}{m^2(\gamma+1)V_-}\right)V_+ + \frac{2(\gamma-1)}{m^2(\gamma+1)}(H_- + qY_-) = 0. \tag{3.2}$$

If we solve this for V_+ we obtain

$$V_+ = \frac{V_-}{\gamma+1}\left(\left(\gamma + \frac{1}{M_-^2}\right) \pm \sqrt{\left(1 - \frac{1}{M_-^2}\right)^2 - \frac{2(\gamma^2-1)qY_-}{u_-^2}}\right), \tag{3.3}$$

where

$$M^2 = \frac{u^2}{c^2} = \frac{u^2}{(\gamma p/\rho)}$$

is the square of the Mach number, and c is the sound speed. Note that if

$$\left(1 - \frac{1}{M_-^2}\right)^2 < \frac{2(\gamma^2-1)qY_-}{u_-^2} \tag{3.4}$$

then (3.1) has no solution. The two values of M^2 where equality holds in (3.4) correspond to exactly one burned state each. These burned states are the Chapman–Jouguet points; see Fig. 1.

For values of M^2 greater than the Chapman–Jouguet detonation value, there are two possible burned states. The flow at the unburned state is supersonic, as is the case for the upstream side of a nonreacting shock wave. At the strong detonation state the flow is subsonic, which corresponds to the downstream side of a nonreacting shock wave. However, at the weak detonation state the flow is supersonic, and this violates the nonreacting shock wave entropy condition.

The effect of the Mach number on the flow pattern for (2.7) at a burned state is as follows. The linearization of (2.7) at a burned state has the following eigenvalues:

$$\begin{aligned}
&\text{(a)} \quad \sigma_1 = 1 - \frac{R}{2c_p} - \frac{RH_+}{m^2V_+^2c_p} = \frac{2(u_+^2 - \gamma p_+/\rho_+)}{2\gamma u_+^2} = \frac{1}{\gamma M_+^2}(M_+^2 - 1) \\
&\text{(b)} \quad \sigma_2 = 1 \\
&\text{(c)} \quad \sigma_{3,4} = \frac{L}{2}\left[1 \pm \sqrt{1 + 4Dk\phi(T_+)u_+^2}\right].
\end{aligned} \tag{3.5}$$

Thus, the strong detonation state has a two-dimensional stable manifold, and the weak detonation state has a one-dimensional stable manifold. From this simple fact it is already clear that although a weak detonation may be possible, a strong detonation is much more likely. The eigenvectors corresponding to $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ are as follows.

$$\begin{aligned}
\sigma_1 &: \quad X_1 = (1,\ 0,\ 0,\ 0), \\
\sigma_2 &: \quad X_2 = ((\gamma - 1),\ m^2\gamma V_+(1 - \sigma_1),\ 0,\ 0), \\
\sigma_3 &: \quad X_3 = \left(\frac{q(\gamma - 1)}{\gamma m^2 V_+(\sigma_3 - \sigma_1)},\ q,\ \sigma_3 - 1,\ \frac{L(\sigma_3 - 1)}{L - \sigma_3}\right), \\
\sigma_4 &: \quad X_4 = \left(\frac{q(\gamma - 1)}{\gamma m^2 V_+(\sigma_4 - \sigma_1)},\ q,\ \sigma_4 - 1,\ \frac{L(\sigma_4 - 1)}{L - \sigma_4}\right).
\end{aligned}$$

From (2.7b) we note that $H_y > 0$ if $H - H_+ + q\epsilon > 0$. On the surface $H - H_+ + q\epsilon = 0$ we have that

$$(H - H_+ + q\epsilon)_y = -\frac{q\lambda}{m^2c_p}r \leq 0. \tag{3.6}$$

Therefore the region where $H_y \geq 0$ is *negatively invariant*, that is, the flow for (2.7) can only exit this region; it cannot enter. On the surface $V_y = 0$ (or $u_y = 0$) we have

$$V_{yy} = \frac{\gamma - 1}{m^2\gamma V_+}H_y. \tag{3.7}$$

Thus, inside the region $H_y \geq 0$, the region $V_y \leq 0$ is negatively invariant. Also the region $H \leq H_+$ is negatively invariant within $H_y \geq 0$.

All of the solutions that we find will lie in the region $H_y \geq 0$, $H \leq H_+$. However, detonation waves that are close to ZND waves will exit the region $V_y \leq 0$, attain a minimum value of V (or u) and maximum values of p and ρ, and then tend to the strong detonation state. To prove the existence of these waves, we need another boundary of the form $V = \textit{constant}$. A natural choice is

$$V = V_0 = \frac{(\gamma - 1)H_-}{(\gamma + 1)m^2V_-}. \tag{3.8}$$

V_0 is the value of V at the end state of a nonreacting shock wave with initial state (V_-, H_-); this is the minimum value of V for a ZND detonation. The part of this

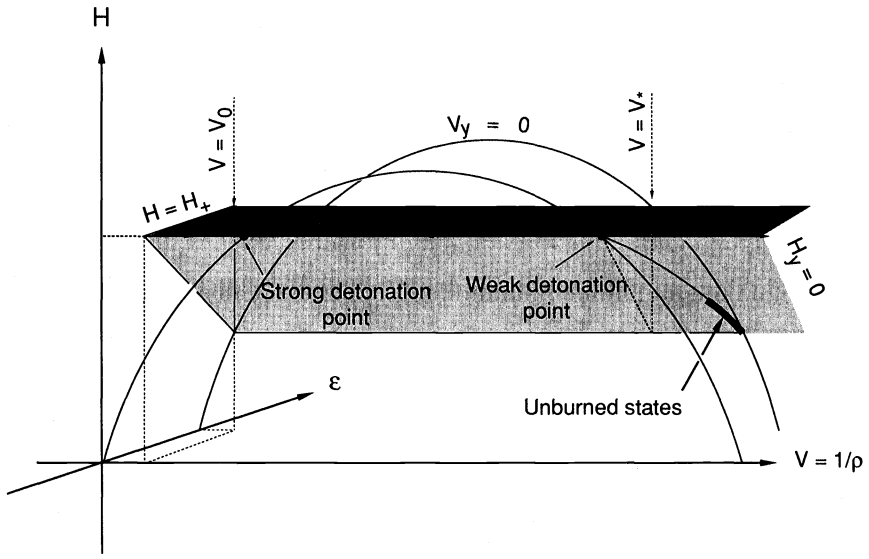


FIG. 2

surface lying within $H_y \geq 0$, $\epsilon < \epsilon_-$, also lies within $V_y \geq 0$. Therefore the region $V \geq V_0$ is positively invariant within the region $H_y \geq 0$, $\epsilon < \epsilon_-$.

Let (V_*, T_*) be the value of (V, T) at the weak detonation state, and let (V^*, T^*) be the value at the strong detonation state. Consider the region W defined by $H_y \geq 0$, $\epsilon \leq \epsilon_-$, $H \leq H_+$, $V_0 \leq V \leq V_*$ (see Fig. 2). The flow for (2.7) enters W through a connected part of the boundary, namely, the union of the surfaces $V = V_*$, $\epsilon = \epsilon_-$, $V = V_0$. The flow leaves W through two components, $H_y = 0$, and $H = H_+$, which are separated by the positively invariant set P defined by $H_y = 0 = H - H_+$, $V_0 \leq V \leq V_*$. The following theorem of Wazewski [Wz1], [Wz2], which we quote from [C], implies that the set of all points in W which eventually flow out of W must also have two components.

DEFINITION 1. Let Γ be a topological space and let $\mathbb{R}$ denote the real numbers. Let a continuous function from $\Gamma \times R \to \Gamma$ be denoted by $(\gamma, t) \to \gamma \cdot t$. This function is called a *flow on* Γ if the following conditions are satisfied for all $\gamma \in \Gamma$ and $s, t \in \mathbb{R}$:

$$\text{(a)} \quad \gamma \cdot 0 = \gamma$$
$$\text{(b)} \quad \gamma \cdot (s + t) = (\gamma \cdot s) \cdot t.$$

If $\Gamma' \subset \Gamma$ and $U \subset \mathbb{R}$, let $\Gamma' \cdot U$ be the set of points $\gamma \cdot t$ such that $\gamma \in \Gamma'$ and $t \in U$.

DEFINITION 2. If $W \subset \Gamma$, let W° be the set of points $\gamma \in W$ such that, for some positive t, $\gamma \cdot t \notin W$. Let W^- be the set of points $\gamma \in W$ such that for any positive t, $(\gamma \cdot (0, t)) \not\subset W$. The set W^- is contained in W and is called the exit set of W. The set W is called a *Wazewski set* if the following conditions are satisfied:

(a) If $\gamma \in W$ and $(\gamma \cdot [0, t]) \subset cl(W)$ then $(\gamma \cdot [0, t]) \subset W$,

(b) W^- is closed relative to W°.

THEOREM. (*Wazewski*). *If W is a Wazewski set then W^- is a strong deformation retract of W° and W° is open relative to W^-.*

The set W that we have described above is a Wazewski set, as is the union of W with the set Q described in the next section. Since W° is homotopic to W, which has two components, W° must also have two components.

What separates these two components is S, the stable manifold for the strong detonation state, because any point which is not in one of these components must stay in W as y tends to infinity. Because H is always increasing within W, this solution must tend to a rest state, namely the strong detonation state. One boundary of this stable manifold is P; another is F, the stable manifold for the weak detonation state.

If there is a connected set in W consisting of points that tend to an unburned state U, as y tends to $-\infty$, and this set intersects both components of the exit set of W, then at least one of these points must be in S. The orbit of such a point is a viscous structure profile for a strong detonation wave.

In the next section we give sufficient conditions for the existence of such a set of points.

4. Some simple estimates. Orbits leaving an unburned state U which lies in $T < T_i$ will stay in the plane $\epsilon = \epsilon_-$ until the surface $T = T_i$ is reached. Consider the curve G given by $H_y \geq 0$, $T = T_i$, $\epsilon = \epsilon_-$, and $V_y \leq 0$. All of the points of G flow towards U as y tends to $-\infty$. The endpoints of G flow out of $H_y \geq 0$, $V_y \leq 0$, immediately. Our strategy is to add a tunnel from W to G so that the exit set of the extended set is still disconnected. The existence of such a tunnel will imply the existence of a viscous structure profile from U to the strong detonation state. The tunnel is:

$$Q = \{(V, H, \epsilon) : V \geq V_*,\ H_y \geq 0,\ V_y \leq 0,\ \epsilon_- \geq \epsilon \geq g(T),\ \text{and } T \geq T_i\}$$

The function g will be chosen so that the flow enters Q through the boundary $\epsilon = g(T)$, and so that $g(T) \geq \epsilon_i$ for $T_i \leq T \leq T_*$, where

$$\epsilon_i = \min\left\{\epsilon \;\middle|\; \begin{array}{l}\text{There exist } H \text{ and } V \text{ such that} \\ H_y(H,V,\epsilon) = V_y(H,V,\epsilon) = \phi(T) = \phi\left(\dfrac{(H - m^2V^2/2)}{c_p}\right) = 0\end{array}\right\}.$$

so that

$$\begin{aligned} q(\epsilon_- - \epsilon_i) &= \left(c_p - \frac{R}{2}\right)(T_i - T_-) \\ &\quad + \frac{u_-^2}{4\gamma^2 M_-^4}\left(1 - \gamma^2 M_-^4 + (\gamma M_-^2 + 1)\sqrt{(\gamma M_-^2 + 1)^2 - 4\gamma M_-^2 \frac{T_i}{T_-}}\right) \end{aligned} \tag{4.1}$$

See Fig. 3. Note that $\epsilon_- - \epsilon_i > 0$ whenever $T_i > T_-$. Also note that the flow enters Q through $\epsilon = \epsilon_-$ and $T = T_i$, while it exits through $H_y = 0$ and $V_y = 0$. The two components of the exit set are separated by the point $H_y = V_y = 0$, $T = T_i$. In order to choose the function g, we note that within Q, we have that

$$\frac{d\epsilon}{dT} = \frac{\epsilon_y}{T_y} = \frac{-\lambda\rho K\left(c_p(T_* - T) + m^2\dfrac{(V_*^2 - V^2)}{2}\right)\phi(T)}{qm^2\left(c_p(T - T_*) - m^2\dfrac{(V - V_*^2)}{2} + q\epsilon - R\left(T - T_*\dfrac{V}{V_*}\right)\right)}. \tag{4.2}$$

The right-hand side of (4.2) is monotone in V within Q. To see that the denominator is monotone in V for $V_* \le V \le V_-$, replace (V_*, T_*) by (V_-, T_-) and differentiate with respect to V, holding T constant. This denominator is also monotone in H as a function of (H, V). In the numerator, we decrease V to V_*. In the denominator, we decrease V, holding T constant, until we reach a value V_1 which satisfies either $H = H_+ - q\epsilon$ or $V_1 = V_*$. In the first case we obtain:

$$\left|\frac{d\epsilon}{dT}\right| \le \frac{\lambda K c_p (T_* - T)\phi(T)}{-m^2 q V_* \left(m^2(V_1^2 - V_1 V_*) - R(T - T_* V_1/V_*)\right)}. \tag{4.3}$$

Then $T = (H_+ - q\epsilon - m^2 V_1^2/2)/c_p$, and we have that:

$$\left|\frac{d\epsilon}{dT}\right| \le \frac{\lambda k c_p^2 (T_* - T)\phi(T)}{R m^2 q^2 V_* (\epsilon - f(V_1))}. \tag{4.4}$$

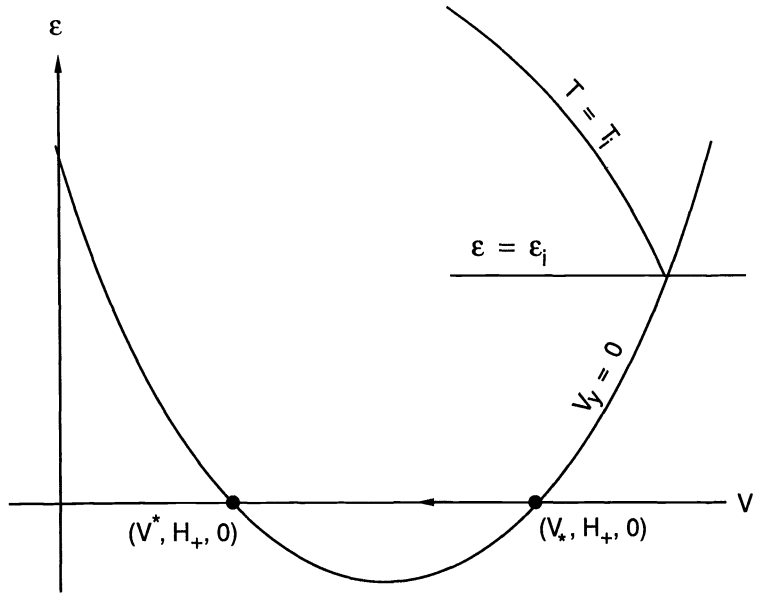


FIG. 3. *The plane* $H_y = H - H_+ + q\epsilon = 0$.

In the second case, we decrease the denominator further by decreasing H while holding $V = V_*$, until we reach $H = H_+ - q\epsilon$. This again yields (4.3), with $V_1 = V_*$ and a smaller value of T in the denominator; however, this value of T equals $(H_+ - q\epsilon - m^2 V_*^2/2)/c_p$. We still obtain (4.4), with $V_1 = V_*$.

Suppose $g(T) \ge \epsilon_i$ for $T_i \le T \le T_*$. Then $\epsilon \ge \epsilon_i$ within Q. Furthermore, $f(V_1) \le \epsilon_i$ within Q (see Fig. 3). Let

$$g'(T) = \frac{\lambda k c_p^2}{R m^2 q^2 V_*} \frac{(T_* - T)\phi(T)}{g(T) - \epsilon_i}, \quad g(T_i) = \epsilon_-. \tag{4.5}$$

Then the flow enters Q through $\epsilon = g(T)$ as long as $g(T) \ge \epsilon_i$. If we solve (4.4), we find that along $\epsilon = g(T)$ we have:

$$(\epsilon - \epsilon_i)^2 = (\epsilon_- - \epsilon_i)^2 - \frac{2\gamma c_p}{(\gamma - 1) m^2 q^2 V_*} \int_{T_i}^{T} \lambda k (T_* - T)\phi(T) dT. \tag{4.6}$$

Since we need to have $\epsilon > \epsilon_i$ for $T_i \le T \le T_*$, we require

$$\epsilon_- > \epsilon_i + \left(\frac{2\gamma c_p}{(\gamma-1)m^2 q^2 V_*} \int_{T_i}^{T_*} \lambda k (T_* - T)\phi(T)\, dT \right)^{1/2}. \tag{4.7}$$

Thus, (4.7) is a sufficient condition for the existence of a strong detonation structure profile. Since both λ and k can depend on T [Wi], it is important to keep them inside the integral.

For a CJ detonation wave, $(V_*, T_*) = (V^*, T^*)$. Condition (4.7) still implies the existence of a structure profile.

As λ tends to zero, with the other parameters held constant, condition (4.7) must be satisfied. In particular, since $\epsilon_- > \epsilon_i$ whenever $T_- < T_i$, we see that the strong and CJ detonations always have viscous structure when λ is sufficiently small. This is the limit of the "small rate parameter" considered by Wood [Wo1].

The region $(V_y < 0) \cap (H_y > 0) \cap (V > V_*) \cap (\epsilon < g(T))$ is negatively invariant, and contains the weak detonation point, and an interval of unburned states, including one corresponding to $\epsilon = \epsilon_i$, $T = T_i$, in its boundary. Since the exit set of this region is homotopic to a circle, and the region itself is contractible, another application of Wazewski's theorem shows that the one-dimensional stable manifold for the weak detonation point must be trapped in this region. Consequently one unburned state with $\epsilon = \epsilon_w$, $T = T_w$, $V = V_w$, which does not satisfy (4.7) must be connected to the weak detonation. Since this weak detonation is a boundary for the stable manifold of the strong detonation point, all unburned states with $\epsilon_- > \epsilon_w$ must be connected to the strong detonation point. As λ tends to zero, (ϵ_w, T_w, V_w) must tend to (ϵ_i, T_i, V_i).

5. A necessary condition. We obtain a necessary condition for the existence of viscous structure using energy estimates similar to those used in work on premixed laminar flames [BNS], [Ma], [Wg]. Observe that

$$H_{yy} = H_y - \frac{\lambda \rho k}{m^2 c_p}(H_+ - H)\phi(T). \tag{5.1}$$

If we multiply (5.1) by 1, H, and then H_y, and integrate each equation from $-\infty$ to $+\infty$, we obtain, using the fact that all derivatives tend to zero as y tends to $\pm\infty$:

$$\begin{aligned}
&\text{(a)} \quad H_+ - H_- = q\epsilon_- = \int_{-\infty}^{\infty} \frac{\lambda\rho k}{m^2 c_p}(H_+ - H)\phi(T)\, dy, \\
&\text{(b)} \quad \int_{-\infty}^{\infty} (H_y)^2\, dy = \int_{-\infty}^{\infty} \frac{\lambda\rho k}{m^2 c_p}(H_+ - H)H\phi(T)\, dy - \frac{H_+^2 - H_-^2}{2}. \\
&\text{(c)} \quad \int_{-\infty}^{\infty} (H_y)^2\, dy = \int_{-\infty}^{\infty} \frac{\lambda\rho k}{m^2 c_p}(H_+ - H)\phi(T)H_y\, dy \\
&\text{(d)} \quad \ge \int_{H_-}^{H_+} \frac{\lambda k}{m^2 c_p V_-}(H_+ - H)\phi\left((H - m^2 V_-^2/2)/c_p\right) dH.
\end{aligned} \tag{5.2}$$

Here we have used the fact that V_- is the maximum value of V on any detonation structure profile. Combining (5.2)(a–c) we find that:

$$\begin{aligned}
(H_+ - H_-)^2 m^2 c_p V_- &= (qY_- m)^2 c_p V_- \\
&\ge \int_{H_-}^{H_+} 2\lambda k (H_+ - H)\phi\left((H - m^2 V_-^2/2)/c_p\right) dH, \\
&= \int_{T_-}^{T_- + qY_-/c_p} 2\lambda k c_p \left(qY_- - c_p(T - T_-)\right)\phi(T)\, dT.
\end{aligned} \tag{5.3}$$

We may obtain a clearer interpretation of (5.3) with a little rearrangement and change of variables:

$$m^2 V_- \geq \int_0^1 2\lambda k c_p (1-\tau)\phi\left(T_- + \frac{qY_-\tau}{c_p}\right) d\tau. \tag{5.3'}$$

Thus, (5.3') constitutes a necessary condition for the existence of a detonation structure profile. Note that if we replace V_- by V_+, then we also obtain a necessary condition for the existence of a deflagration structure profile. Also, (5.3') is satisfied whenever $m^2 V_- \geq \lambda k \phi(T_- + qY_-/c_p)/c_p$.

6. Arbitrary Lewis numbers. When the Lewis number, $L = \lambda/\rho D c_p$, is not identically 1, then we must work with all four equations of (2.7). Consequently the region R must be extended to a region in $\mathbb{R}^4$. We require, therefore, additional boundaries for R, that is, upper and lower bounds for Y. It is interesting that we can obtain natural bounds for Y in terms of H, similar to the bounds that have been obtained for premixed laminar flames.

LEMMA. *Let*

$$\Lambda_* = \inf(L, 1) \tag{6.1}$$

$$\Lambda^* = \sup(L, 1)$$

Then

$$\Lambda_*(H_+ - H) \leq qY \leq \inf(q\epsilon, \Lambda^*(H_+ - H)) \tag{6.2}$$

defines a negatively invariant region for (2.7).

Proof. We proceed exactly as in [Wg], using (2.8):

$$\begin{aligned}\frac{d}{dy}\left(H - H_+ + \frac{qY}{\Lambda_*}\right) &= (H - H_+)_y + \frac{q}{\Lambda_*} Y_y \\ &\leq (H - H_+)_y + \frac{q\rho D c_p}{\lambda} Y_y \\ &= H - H_+ + qY \\ &\leq H - H_+ + \frac{q}{\Lambda_*} Y.\end{aligned}$$

This last quantity is zero on the boundary of the region defined by

$$H - H_+ + \frac{q}{\Lambda_*} Y \geq 0.$$

Consequently this region is negatively invariant for (2.7). Similarly, the region defined by

$$H - H_+ + \frac{q}{\Lambda^*} Y \leq 0$$

is negatively invariant, within the region $Y \leq \epsilon$. This last condition is required to ensure that $Y_y \leq 0$. The region $Y \leq \epsilon$ is negatively invariant because on $Y = \epsilon$,

$$\begin{aligned}\frac{d}{dy}(\epsilon - Y) &= \epsilon_y - Y_y \\ &= \epsilon_y \leq 0.\end{aligned}$$

□

The topological argument is a little more sophisticated than the one given in §3. The region W is defined as before, with the additional inequality (6.2). The exit set for W is now connected; using the lemma we see that it is a union of four parts: $\Sigma_{out} = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3 \cup \Sigma_4$, where

$$\Sigma_1 = \{(V,\ H,\ \epsilon,\ Y) \mid H = H_+,\ Y = 0,\ 0 < \epsilon \le \epsilon_-,\ V_0 < V < V_*\}$$

$$\Sigma_2 = \left\{(V,\ H,\ \epsilon,\ Y) \,\middle|\, \begin{array}{l} qY = \min(q\epsilon,\ \Lambda^*(H_+ - H)),\ 0 < \epsilon \le \epsilon_-, \\ V_0 < V < V_*,\ 0 \le H_+ - H \le q\epsilon \end{array}\right\}$$

$$\Sigma_3 = \{(V,\ H,\ \epsilon,\ Y) \mid H_+ - H = q\epsilon,\ \Lambda_*\epsilon \le Y \le \epsilon,\ 0 < \epsilon \le \epsilon_-,\ V_0 < V < V_*\}$$

$$\Sigma_4 = \left\{(V,\ H,\ \epsilon,\ Y) \,\middle|\, \begin{array}{l} qY = \Lambda_*(H_+ - H),\ 0 < \epsilon \le \epsilon_-, \\ V_0 < V < V_*,\ 0 \le H_+ - H \le q\epsilon \end{array}\right\}.$$

Note, however, that Σ_{out} is homotopic to a circle. The circle may be visualized as a path from Σ_1 to Σ_2 to Σ_3 to Σ_4 and back to Σ_1. This path cannot be contracted to a point within Σ_{out} because Σ_{out} does not contain the positively invariant line segment

$$P = \{(V,\ H,\ \epsilon,\ Y) : qY = H_+ - H = q\epsilon = 0, V_0 \le V \le V_*\}.$$

Again, Wazewski's principle implies that the set W_{out} of all points in W which eventually flow out of W must also be homotopic to a circle. The curve G of §4, given by $H_y \ge 0$, $T = T_i$, $\epsilon = \epsilon_-$, $V_y \le 0$, is now extended, via (6.2), to a set G^* which is homeomorphic to a disk. The tunnel Q is also extended via (6.2). The estimates of §4 now show that the exit set of $W \cup Q$ is also homotopic to a circle. The boundary of G^* is homeomorphic to a circle and intersects the exit set of $W \cup Q$ with nonzero degree. Since a disk is not homotopic to a circle, at least one of the points of G^* must not exit W. This point must in fact tend to the strong detonation burned state. The sufficient condition (4.7) is now

$$\epsilon_- > \epsilon_i + \left(\frac{2\gamma c_p}{(\gamma-1)m^2q^2V_*}\int_{T_i}^{T_*} \lambda k\Lambda^*(T_* - T)\phi(T)dT\right)^{1/2}. \tag{6.3}$$

7. Behavior. We have shown that strong or CJ detonation structure profiles exist if (4.7) is satisfied. From this we can see that the existence of these profiles depends on the values of λ, ϕ, and k, relative to m, q, and γ, between the unburned state and the weak detonation point. We will now show that the behavior of the solution depends strongly on the values of these parameters near the strong detonation point.

At the strong detonation point the linearization of (2.7) has two negative eigenvalues:

$$\text{(a)}\quad \sigma_1 = \frac{1}{\gamma}\left(1 - M^{*-2}\right)$$

$$\text{(b)}\quad \sigma_3 = \frac{L}{2}\left(1 - \sqrt{1 + \frac{4Dk\phi(T^*)}{u^{*2}}}\right). \tag{7.1}$$

The relative sizes of σ_1 and σ_3 determine the node structure of the flow for (2.7) at the strong detonation point. If $Dk\phi(T^*)/u^{*2}$ is very small, or if L is very small (which

would be physically unusual), so that $0 > \sigma_3 > \sigma_1$, then the flow has a node tangent to the eigenvector

$$X_3 = \left[\frac{q(\gamma-1)}{\gamma m^2 V^*(\sigma_3-\sigma_1)},\ q,\ \sigma_3-1,\ \frac{L(\sigma_3-1)}{L-\sigma_3}\right].$$

Thus, all but one of the structure profiles approaches the strong detonation state tangent to X_3. The one orbit that approaches tangent to X_1 is the purely nonreacting shock profile which connects the weak detonation state to the strong detonation state. The $(V,\ H,\ \epsilon)$ components of X_3 have the signs $(+,\ +,\ -)$. Thus, at the end of the wave, V and H are increasing and ϵ, which must be monotone, is decreasing. Since $V^* < V_-$, V cannot be monotone throughout the entire solution. Once the solution leaves the region $V_y < 0$, it cannot reenter, because H is monotone increasing. Hence V attains its minimum value at a single point of the solution. This minimum value must be greater than V_0, the minimum value for a ZND detonation.

The pressure must also attain an extremum in this case:

$$\begin{aligned}\frac{dp}{dy} &= \frac{R}{V}\frac{dT}{dy} - \frac{RT}{V^2}\frac{dV}{dy}\\ &= \frac{R}{c_p}\left(\frac{H_y}{V} - m^2\left(1+\frac{c_pT}{u^2}V_y\right)\right).\end{aligned}$$

Since the flow is tangent to X_3 at the strong detonation state, we have, as the solution approaches that point,

$$\begin{aligned}\frac{dp}{dy} &= \frac{RH_y}{c_pV^*}\left(1-\frac{\gamma-1+M^{*-2}}{\gamma(\sigma_3-\sigma_1)}\right)\\ &= \frac{RH_y}{c_pV^*(\sigma_3-\sigma_1)}\left(\frac{L}{2}\left(1-\sqrt{1+\frac{4Dk\phi(T^*)}{u^{*2}}}\right)-1\right)\\ &< 0.\end{aligned}$$

Since p is increasing when $V_y < 0$, p must attain a maximum in the interior of the wave.

The temperature behaves differently. Near the strong detonation state, we have

$$\begin{aligned}\frac{dT}{dy} &= \frac{1}{c_p}\left(H_y - m^2VV_y\right)\\ &= \frac{1}{c_p}\left(\frac{\gamma m^2V^*(\sigma_3-\sigma_1)}{\gamma-1} - m^2V^*\right)V_y\\ &= \frac{m^2V^*}{c_p(\gamma-1)}\left(\frac{\gamma L}{2}\left(1-\sqrt{1+\frac{4Dk\phi(T^*)}{u^{*2}}}\right)+M^{*-2}-\gamma\right)V_y.\end{aligned}$$

Thus, if M^{*-2} is less than γ, we see that T decreases near the burned state. Consequently T must attain a maximum value in the interior of the wave. This is consistent with the behavior of the ZND wave corresponding to this case; see [Wi, pp. 194–197].

If M^{*-2} is greater than γ , then T decreases near the burned state if $\sigma_3 - \sigma_1$ is sufficiently small, but positive. In the inviscid limit for this case, as D, λ, and μ tend to zero, σ_3 tends to zero and T must increase near the burned state. It is a natural

conjecture that T is monotone throughout the wave in this case. We say, therefore, that the flow is *strongly subsonic* near the burned state if $\gamma M^{*2} < 1$. If this condition is satisfied, then the maximum value of the temperature on the surface $V_y = 0$ (the Rayleigh line) occurs at a value of H higher than H_+, for on the surface $V_y = 0$,

$$\frac{dT}{dV} = \frac{T}{V}\left(1 - \gamma M^2\right).$$

and

$$\frac{dH}{dV} = \frac{Vm^2}{\gamma - 1}\left(1 - M^2\right),$$

whereas

$$\frac{d\left(M^2\right)}{dV} = \frac{M^2}{V}\left(1 + \gamma M^2\right) > 0.$$

(See [Wi]). As a consequence, the corresponding ZND wave, which follows the intersection of the surface $V_y = 0$ and the plane $H_y = 0$, must have a monotone temperature whenever the burned state is strongly subsonic.

This phenomenon may be explained by noting that M_* depends on the total heat released per unit mass, qY_-, and increases as qY_- increases with other parameters, particularly m, held constant. Thus the strong detonation burned state is strongly subsonic if the heat release is too weak, relative to the strength of the wave, to create a temperature spike. In experiments (see [Wi, §6.2.1] and references cited therein), strong detonation waves are observed principally when a piston, or other external force, is used to overdrive the wave; in this respect strong detonation waves resemble inert shock waves. Detonation waves that are not overdriven will decay to a Chapman–Jouguet detonation. Chapman–Jouguet detonations may be thought of as reacting shock waves that are driven by the reaction with no external force. Thus these waves are "pure" reacting shock waves, and qY_- is a parameter between inert shock waves with monotone temperature profiles and Chapman–Jouguet waves with temperature spikes.

As M^* tends to 1, σ_1 tends to zero, which leads us to the next case. If $Dk\phi(T^*)/u^{*2}$ is very large, or if L is very large, or if M_* is very close to 1, so that $0 > \sigma_1 > \sigma_3$, then the flow forms a node tangent to the eigenvector

$$X_1 = (1,\ 0,\ 0,\ 0).$$

In this case the solution may behave in an unusual manner. Detonations with Y_- close to ϵ_i, (T_- close to T_i) will follow the trajectory of the weak detonation, and then turn near the weak detonation burned state, and approach the strong detonation point along the trajectory of the inert shock profile. Thus these structure profiles look like a weak detonation followed by an inert shock wave. See Fig. 4, where a heuristic picture of the flow in the stable manifold is presented.[1] Similar observations have been made in [Wo2], [FD], [LL], and most recently for numerical calculations and for a simpler model in [CMR], where it was noted that these pathological waves are actually numerically stable as solutions of the time-dependent reacting compressible

[1]Note that we have not proved that the stable manifold has no folds, so that V and e are global coordinates for the stable manifold, or that there is a monotone relationship between Y_- and the solution curves, such as depicted in Fig. 4. For a proof of such monotonicity for premixed laminar flames with $L > 1$, see [Ma].

Navier–Stokes equations. These authors also observed (numerically) an interesting phenomenon, namely that if the turning point of these solutions is sufficiently close to the weak detonation state, that is, if the spatial separation between the weak detonation and the inert shock wave is sufficiently large, then a bifurcation occurs wherein these two waves decouple and the inert shock moves slower than the weak detonation as predicted in [Wo2]. In this sense, a weak detonation can be observed. Weak detonations are also observed experimentally [FD], as a consequence of very complicated chemistry, change in equation of state, and endothermic or reversible reactions. A more significant point, made in [CMR], is that since $k\phi(T^*)/u^{*2}$ can be significantly large, it is important, in making machine calculations, to use an approximation scheme which does not add too much artificial diffusion, because this can radically change the character of the solution.

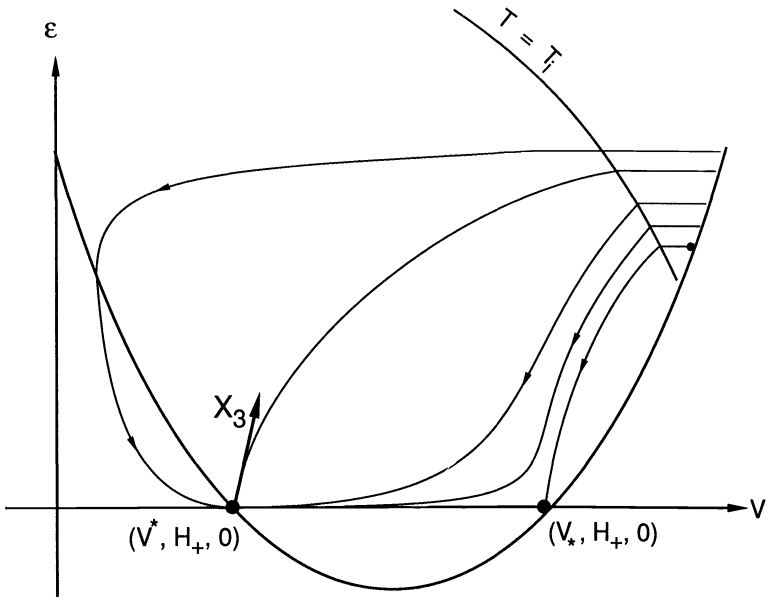


FIG. 4. *The flow in the stable manifold when* $0 > \sigma_1 > \sigma_3$.

For larger values of Y_-, there will be one wave which approaches the strong detonation state tangent to X_3, and others which approach tangent to X_1 but with V increasing. For the singular wave tangent to X_3 all quantities will be monotone, but for the others the pressure and density will attain maximum values. Since $H = c_pT + (m/\rho)^2/2$ is constant along X_1, the existence of a density peak in these waves implies the existence of a temperature peak.

For the CJ detonation, the burned state Mach number is 1, so that the above remarks apply to the behavior of this wave. All but one of the CJ detonation structure profiles must approach tangent to the eigenvector $\pm(1,0,0,0)$. However, solutions cannot approach along $+(1,0,0,0)$ (the right side) because $V_y > 0$ there. Consequently there is one monotone CJ structure profile (presumably the one with minimum heat release) and the rest have peaks in pressure, density, and temperature.

8. The ZND limit. We have noted that as λ tends to zero (or λ and D, with L constant), then the sufficient condition (4.7), or (6.3), must be satisfied. The structure

profiles satisfy natural a priori bounds, namely

$$V_0 \le V \le V_-,$$
$$H_- \le H \le H_+,$$
$$0 \le \Lambda_* \epsilon \le Y \le \epsilon \le Y - .$$

Thus, for any sequence $\lambda_n \to 0$, there is a corresponding sequence $(V_n,\ H_n,\ \epsilon_n,\ Y_n)$ of structure profiles with fixed end states, and which are uniformly bounded. Since H, ϵ, and Y are monotone, and V has at most one minimum, these profiles are also uniformly bounded in total variation. By Helly's theorem, some subsequence of this sequence must converge to a limit in L^1_{loc}. Taking a further subsequence we obtain convergence almost everywhere. The limit function is therefore a weak solution to (1.1a–c, d′). If $T_- < T_i$ then the limit is a ZND strong or CJ detonation. If $T_- = T_i$ then the limit is a continuous weak detonation which follows the curve $(H_y = 0) \cap (V_y = 0) \cap (\epsilon = Y)$ from $(V_i,\ H_i,\ \epsilon_i,\ Y_i)$ to $(V_*,\ H_*,\ 0,\ 0)$.

9. The second law of thermodynamics. In several shock structure problems, and particularly in MHD [CS1], [CS2], [Ge], the entropy flux has played an important role. In MHD the structure equations take the form

$$\frac{du}{dx} = B \nabla P(u)$$

where B is a positive diagonal matrix and P is the entropy flux. The gradient-like structure that this gives to the problem is essential to our understanding of the solution to this very complicated system. In other areas it is useful to postulate that the entropy flux must be monotone [HaSe1],[HaSe2]; the inequality expressing this monotonicity is called the *Clausius–Duhem inequality.* We show here, assuming only that the reaction is exothermic and irreversible ($r(\rho, Y, T) \ge 0$), that the entropy flux is monotone along a viscous structure profile for a plane steady detonation or deflagration wave. We have not used this fact in the above discussion, although it would have proved useful if we had not already known that H must be monotone.

The entropy flux is

$$P = mS - \lambda T_x / T.$$

Using Gibb's law: $TdS = de + pdV - qdY$ [Wi], we find that

$$\frac{dP}{dx} = \frac{m}{T}\left(\frac{de}{dx} + p\frac{dV}{dx} - q\frac{dY}{dx}\right) + \frac{\lambda T_x^2}{T^2} - \frac{(\lambda T_x)_x}{T}.$$

Using (1.2), we have

$$\begin{aligned}\frac{dP}{dx} &= \frac{\lambda T_x^2}{T^2} - \frac{1}{T}\left(\left[\left[\rho\left(u^2/2 + e(\rho,T,Y)\right) + p(\rho,T)\right]u\right]_x - (\mu u u_x)_x - (q\rho D Y_x)_x\right) \\ &\quad + \frac{m}{T}\left(\frac{de}{dx} + p\frac{dV}{dx} - q\frac{dY}{dx}\right) \\ &= \frac{\lambda T_x^2}{T^2} + \frac{\mu u_x^2}{T} + \frac{qr}{T} \ge 0.\end{aligned}$$

Note that dP/dx only vanishes at rest points of (1.2). However, (1.2) is not gradient-like with respect to P, because

$$dP = -\frac{\lambda T_x}{T} dT + \frac{q d\epsilon}{T} - \frac{\mu u_x}{T} du;$$

since q/T does not vanish, neither does the gradient of P. This may be an anomaly which is due to the simple representations of the chemistry and the reaction rate which are used here.

Acknowledgments. The author would like to thank Professor Andrew Majda for his encouragement and support, the National Science Foundation and the Air Force Office of Scientific Research for their financial support, and Steffen Heinze for pointing out several typographical mistakes.

REFERENCES

[BNS] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Travelling wave solutions to combustion models and their singular limits.*, SIAM J. Math. Anal., 16 (1985), pp. 1207-1242.

[CMR] P. COLELLA, A. MAJDA, AND V. ROYTBURD, *Theoretical and numerical structure for reacting shock waves*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 1059-1080.

[C] C. C. CONLEY, *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conference Series in Mathematics, 38, American Mathematical Society, Providence, RI, 1978.

[CS1] C. C. CONLEY AND J. A. SMOLLER, *On the structure of magnetohydrodynamic shock waves*, Comm. Pure Appl. Math., 27 (1974), pp. 367-375.

[CS2] ———, *On the structure of magnetohydrodynamic shock waves* II, J. Math. Pures Appl., 54 (1975), pp. 429-444.

[D] W. DÖRING, *The detonation process in gases*, Ann. Physik, 43 (1943), p. 421.

[FD] W. FICKETT AND W. DAVIS, *Detonation*, University of California Press, Berkeley, Los Angeles, 1979.

[Ga] R. A. GARDNER, *On the detonation of a combustible gas*, Trans. Amer. Math. Soc., 277 (1983), pp. 431-468.

[Ge] P. GERMAIN, *Shock waves, jump relations, and structure*, Adv. in Appl Mech., 12 (1972), pp. 131-194.

[Gi] D. GILBARG, *The existence and limit behavior of the one-dimensional shock layer*, Amer. J. Math., 73 (1951), pp. 256-275.

[HaSe1] R. HAGAN AND J. SERRIN, *Dynamic changes of phase in a van der Waals fluid*, in New Perspectives in Thermodynamics, J. Serrin, ed., Springer-Verlag, Berlin, New York, 1986, pp. 241-260.

[HaSe2] ———, *One-dimensional shock layers in Korteweg fluids*, in Phase Transformations and Material Instabilities in Solids, M. Gurtin, ed., Academic Press, New York, 1984.

[HC] J. O. HIRSCHFELDER AND C. C. CURTISS, *Theory of Detonations. I. Irreversible unimolecular reaction*, J. Chem. Phys., 28 (1958), pp. 1130-1147.

[HoSt] P. HOLMES AND D. S. STEWART, *The existence of one dimensional steady detonation waves in a simple model problem*, Stud. Appl. Math., 66 (1982), pp. 121-143.

[LL] G. C. LU AND G. S. S. LUDFORD, *Asymptotic analysis of plane steady detonations*, SIAM J. Appl. Math., 42 (1982), pp. 625-635.

[Ma] M. MARION, *Qualitative properties of a nonlinear system for laminar flames without ignition temperature*, Nonlinear Anal., 9 (1985), pp. 1269-1292.

[N1] J. VON NEUMANN, *Theory of detonation waves*, Progress Report 238, Office of Scientific Research and Development, Report 549 (1942); Ballistic Research Lab. File X-122. Also found in [N3].

[N2] ———, *Theory of shock waves*, Division 8, N.D.R.C., (1943) Office of Scientific Research and Development, Report 1140. Also found in [N3].

[N3] ———, *Collected Works*, Pergamon Press, Elmsford, NY, 1963, pp. 178-218.

[Wg] D. H. WAGNER, *Premixed laminar flames as travelling waves*, in Reacting Flows: Combustion and Chemical Reactors, G. S. S. Ludford, ed., Lecture Notes in Appl. Math. 24, American Mathematical Society, Providence, RI, 1986, pp. 229-237.

[Wz1] T. WAZEWSKI, *Sur un principe topologique de l'examen de l'allure asymptotiques des intégrals des équations differentielles*, Ann. Soc. Polon. Math., 20 (1947), pp. 279-313.

[Wz2] ———, *Sur une méthode topologique de l'examen de l'allure asymptotiques des intégrals des équations differentielles*, in Proc. Internat. Congr. Math., Amsterdam, 1954, vol. 3, Noordhoff, Groningen, the Netherlands, 1956, pp. 132-139.

[Wi] F. A. WILLIAMS, *Combustion Theory*, Benjamin/Cummings, Menlo Park, CA, 1985.

[Wo1] W. W. WOOD, *Existence of detonations for small values of the rate parameter*, Phys. Fluids, 4 (1961), pp. 46-60.

[Wo2] W. W. WOOD, *Existence of detonations for large values of the rate parameter*, Phys. Fluids, 6 (1963), pp. 1081-1090.

[Z] YA. B. ZEL'DOVICH, *Theory of the propagation of detonations in gaseous systems*, Soviet Phys. JETP, 10 (1942), pp. 542ff.

[ZK] YA. B. ZEL'DOVICH AND A. S. KOMPANEETS, *Theory of Detonation*, Academic Press, New York, 1960.

[ZR] YA. B. ZEL'DOVICH AND YU. P. RAIZER, *Physics of shock waves and high-temperature hydrodynamic phenomena*, Academic Press, New York, London, 1966.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1055-1068, September 1989

002

BERNSTEIN FUNCTIONS AND THE DIRICHLET PROBLEM*

ALAN R. ELCRAT† AND KIRK E. LANCASTER†

Abstract. For a nonconvex, symmetric quadrilateral, the nonparametric minimal surface arising from an associated Dirichlet problem can be described in terms of the Weierstrass representation and the stereographic projection of its Gauss map. The Bernstein function—which arises by truncation of the re-entrant corner by a concave arc and by requiring the normal vector to be horizontal there—has the same Gauss map image. This leads to a Riemann-Hilbert problem that can be solved and leads to the existence of this surface.

Key words. nonparametric minimal surface, Riemann-Hilbert problem, Bernstein function

AMS(MOS) subject classifications. 35J65, 53A10

Introduction. If Ω is a nonconvex, planar domain and ϕ is a continuous function, defined on $\partial\Omega$, for which the Dirichlet problem for the minimal surface equation does not have a continuous solution, it is natural to ask about the behavior of generalized, "variational" solutions near boundary points where the solution is discontinuous and to determine the geometric nature of the graph. In the particular case of a symmetric, nonconvex quadrilateral with boundary values that are zero on the outer edges and which increase linearly to a positive value on the edges adjacent to the re-entrant corner, these questions have been asked by Finn [F1] and Nitsche [N] and have been studied recently by Elcrat and Lancaster [EL]. If the latter results are combined with the subsequent work of Lancaster [L1], we can assert that this problem is well understood. In fact, an experimental soap film realization of the solution is shown in [EL]. The domain in this case is regular for the minimal surface equation at all but one of its boundary points, and the next step in complexity is a domain whose boundary consists of two arcs, one convex and one concave. In particular we can consider a truncation of the re-entrant quadrilateral by a concave arc. The critical first step in understanding this new problem is the construction of a Bernstein function, a solution of the minimal surface equation whose normal derivative is infinite there. This surface serves as an upper bound for the trace of generalized solutions on the concave arc and can be used as a comparison function for other domains.

In this work we reconsider the solution of the re-entrant quadrilateral problem and, in the process, obtain an existence theorem for a Bernstein function in the truncated quadrilateral. The approach that we use is based on the intuitive observation that for these two surfaces the Gauss maps have the same image. By a stereographic projection we can consider this image as a domain in the complex plane and if we introduce the Weierstrass representation of minimal surfaces in terms of analytic functions f and g, using *this* domain for the complex parameter, the second function g reduces to the identity. Furthermore, f satisfies explicit linear boundary conditions given by properties of the normal vector of the surface sought. This means that our problem reduces to a Riemann-Hilbert problem for the analytic function f. The remarkable fact is that the solution of the re-entrant quadrilateral problem and the Bernstein function satisfy conditions that differ only by an inhomogeneous term in one of the boundary conditions. It turns out that, using results for Riemann-Hilbert problems with discontinuous coefficients [M], we can deduce the existence of a solution of the inhomogeneous

* Received by the editors September 28, 1987; accepted for publication (in revised form) October 31, 1988.
† Department of Mathematics and Statistics, Wichita State University, Wichita, Kansas 67208. The first author's research was partially supported by Air Force Office of Scientific Research grant 86-0274.

Riemann-Hilbert problem from existence for the homogeneous problem. We must prove, in order to complete the chain of deductions, that the solution of the re-entrant quadrilateral problem is a surface whose Gauss map has the required properties and that the analytic function f solving the inhomogeneous problem leads to a surface of the required type. These tasks demand some effort, which accounts for most of the text of the paper.

The significance of this work derives not only from proving the existence of a useful comparison function, but from its construction using a geometric transformation that changes a nonlinear problem into a linear one, which can be solved by giving the solution of a boundary value problem for an analytic function. This idea currently relates to a variety of problems in applied mathematics—for example, numerical conformal mapping [H], [T] and free boundary problems in fluid mechanics [M], [ET]. Although the idea of reconstructing a surface from its Gauss map has been known for a long time, we are not aware of many examples in which an explicit construction can be given of a surface whose boundary geometry is prescribed. Furthermore, the construction of a new minimal surface from a known one using our procedure is new.

1. The re-entrant quadrilateral problem. In this section, we will specify the geometry, set the notation, and state some required theorems from previous work. Let $\alpha, \beta, \gamma \in (0, \pi/2)$ with $\alpha+\gamma<\pi/2$. Set $A_0=(0, -\sin(\gamma)/\sin(\alpha))$, $A_1=(-\sin(\alpha+\gamma), \cos(\alpha+\gamma))$, $A_2=A_3=(0,0)$, and $A_4=(\sin(\alpha+\gamma), \cos(\alpha+\gamma))$ (see Fig. 1). Let Ω be the region bounded by the nonconvex symmetric quadrilateral $A_3A_4A_0A_1A_2$. Let $\phi \in C^0(\partial\Omega)$ equal zero on $A_1A_0A_4$, increase linearly from 0 at A_1 to $\tan(\beta)$ at A_2, and increase linearly from 0 at A_4 to $\tan(\beta)$ at A_3.

Set

$$\Gamma_1=\{(x, y, 0)\,|\,(x, y)\in A_0A_1\},$$
$$\Gamma_2=\{(x, y, \phi(x, y))\,|\,(x, y)\in A_1A_2\},$$
$$\Gamma_4=\{(x, y, \phi(x, y))\,|\,(x, y)\in A_3A_4\},$$
$$\Gamma_5=\{(x, y, 0)\,|\,(x, y)\in A_4A_0\},$$

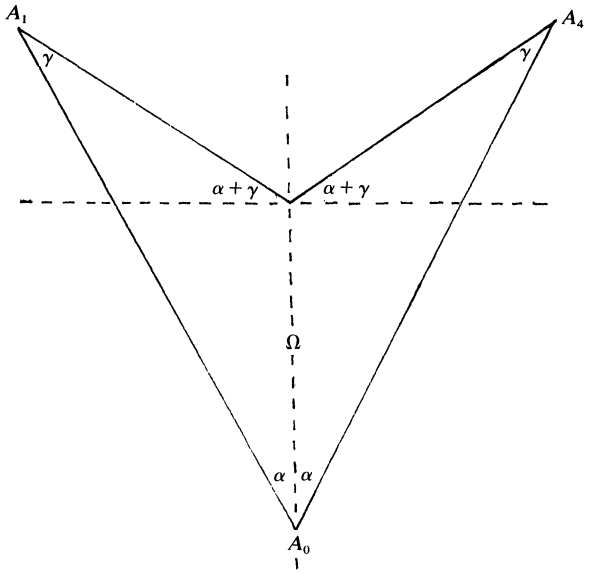


FIG. 1

where, for example, A_0A_1 is the closed line segment between A_0 and A_1. Define Γ to be $\Gamma_1 \cup \Gamma_2 \cup \Gamma_4 \cup \Gamma_5$.

Set

$$T_1 = \{(u, v) \mid v = \tan(\alpha)u\},$$
$$T_2 = \{(u, v) \mid (u+u_0)^2 + (v-v_0)^2 = r^2\},$$
$$T_3 = \{(u, v) \mid u^2 + v^2 = 1\},$$
$$T_4 = \{(u, v) \mid (u-u_0)^2 + (v-v_0)^2 = r^2\},$$
$$T_5 = \{(u, v) \mid v = -\tan(\alpha)u\},$$

where $(u_0, v_0) = \cot(\beta)(\sin(\alpha+\gamma), \cos(\alpha+\gamma))$, and $r = \csc(\beta)$.

Let

$$H_1 = \{(x, y, z) \mid y = \tan(\alpha)x\},$$
$$H_2 = \{(x, y, z) \mid -\cos(\beta)\sin(\alpha+\gamma)x + \cos(\beta)\cos(\alpha+\gamma)y + \sin(\beta)z = 0\},$$
$$H_3 = \{(x, y, z) \mid z = 0\},$$

and

$$H_4 = \{(x, y, z) \mid \cos(\beta)\sin(\alpha+\gamma)x + \cos(\beta)\cos(\alpha+\gamma)y + \sin(\beta)z = 0\},$$
$$H_5 = \{(x, y, z) \mid y = -\tan(\alpha)x\}.$$

Note that T_k is the stereographic projection of the great circle $H_k \cap S^2$, for $k = 1, \cdots, 5$, and that H_k is orthogonal to Γ_k, for $k = 1, 2, 4, 5$. The stereographic projection of any unit vector orthogonal to Γ_k or to the z-axis will be in T_k or T_3 ($k = 1, 2, 4, 5$).

Let w_0 be the origin and let w_k be the point of intersection of $T_k \cap T_{k+1}$ that lies in the upper half-plane for $k = 1, \cdots, 4$. Let σ_k be the shorter (closed) arc of T_k between w_{k-1} and w_k, $k = 1, \cdots, 5$, where $w_5 = w_0$. Set $\sigma = \sigma_1 \cup \cdots \cup \sigma_5$ and define D ($= D(\alpha, \beta, \gamma)$) to be the open region in the plane bounded by σ. Note that D is contained in the open unit disc B and $\sigma_3 = \partial D \cap \partial B$ (see Fig. 2).

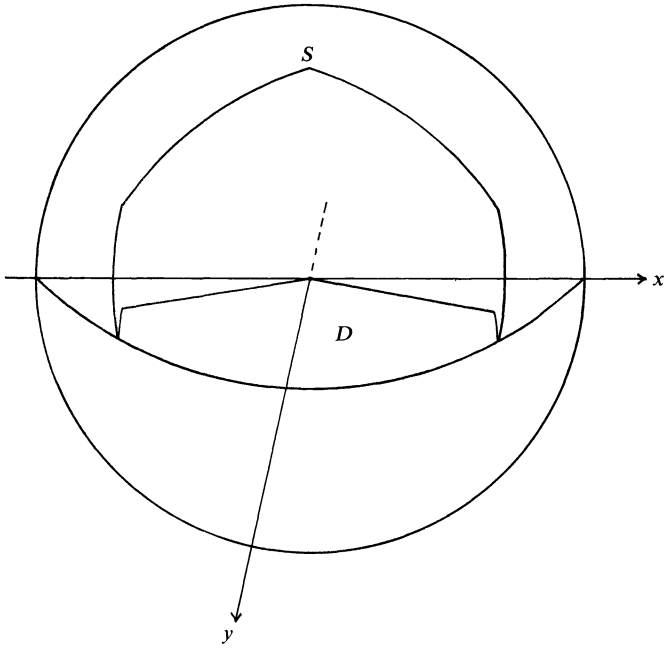


FIG. 2

Set $B^+ = \{(u, v) \in B \mid v > 0\}$. We will consider D, B^+, and B as subsets of $\mathbb{C}$ when this is convenient. When doing so, we will let $w = u + iv$ represent a point of D and $\omega = u + iv$ represent a point of B^+ or B. D will be shown to be the image of the Gauss maps for the surfaces discussed in this paper.

Let $F \in BV(\Omega)$ minimize

$$J(v) = \int\int_{\Omega} \sqrt{1 + |Dv|^2} + \int_{\partial\Omega} |v - \phi|$$

over $v \in BV(\Omega)$. Then $F \in C^2(\Omega) \cap C^0(\bar{\Omega} \backslash \{A_2\})$ and $F = \phi$ on $\partial\Omega \backslash \{A_2\}$. Define S_0 to be the graph of F over Ω and S to be the closure of S_0. The graph of F is the re-entrant quadrilateral surface discussed in [EL] (see the Introduction). From [L1] and [L2] we obtain Proposition 1.

PROPOSITION 1. *There exists* $Y \in C^2(B^+: \mathfrak{R}^3) \cap C^0(\overline{B^+}: \mathfrak{R}^3)$ *such that* Y *is conformal and has harmonic components, and* Y *maps* B^+ *homeomorphically onto* S_0, $\partial' B^+$ $(= \partial B^+ \cap \partial B)$ *strictly monotonically onto* Γ, *and* $\partial'' B^+$ $(= \partial B^+ \cap B)$ *into the z-axis. Furthermore,* $Y(-1, 0) = Y(1, 0) = (A_2, \phi(A_2))$, $Y(-u, v) = \operatorname{diag}(-1, 1, 1) Y(u, v)$ *for* $(u, v) \in B^+$, Y *has a branchpoint at* $(0, 0)$, Y *maps* $\{(u, v) \in \partial B^+ \mid u \geqq 0, v \geqq 0\}$ *onto* $\Gamma_4 \cup \Gamma_5$, *and* Y *extends by reflection across* $\partial'' B^+$ *as a parametric minimal surface.*

It will be useful for us to transplant this parametrization of S to another domain. Let us define

$$\frac{d}{dw} = \frac{1}{2}\left(\frac{\partial}{\partial u} - i\frac{\partial}{\partial v}\right), \qquad \frac{d}{d\omega} = \frac{1}{2}\left(\frac{\partial}{\partial u} - i\frac{\partial}{\partial v}\right).$$

Then $Y_\omega(0) = 0$. Let h_1 be the conformal map from D onto B^+ sending w_0 to i, w_2 to -1, and w_3 to 1 and let h_2 be the conformal map of B onto B^+ sending i to i, -1 to -1, and 1 to 1. Define $h: D \to B$ by $h = h_2^{-1} \circ h_1$. Note that h_1 extends analytically across $\partial D \backslash \{w_0, \cdots, w_4\}$ and has nonvanishing derivative on $\bar{B} \backslash \{w_0, \cdots, w_4\}$ (e.g., [GL]). Similarly, h_2 extends analytically across $B \backslash \{-1, 1\}$ and has nonvanishing derivative on $\bar{B} \backslash \{-1, 1\}$. The asymptotic behavior of h_1 and h_2 at corners is known (i.e., [H, p. 359]). Set

$$X(w) = Y(h_1(w)) \quad \text{for } w \in \bar{D},$$

$$\tilde{X}(\omega) = Y(h_2(\omega)) \quad \text{for } \omega \in \bar{B}.$$

Let $X(w) = (x(w), y(w), z(w))$. The symbols X, $x(w)$, $y(w)$, and $z(w)$ will represent the parametric minimal surface defined above and its components throughout this article.

For $k = 0, 1, 4$, define $\omega_k \in \partial B$ so that $\tilde{X}(\omega_k) = (A_k, \phi(A_k))$ and define $\omega_2 = -1$ and $\omega_3 = 1$. Note $\omega_0 = i$ by symmetry. Let s_k be the (closed) arc on ∂B between ω_{k-1} and ω_k that does not contain any of the other ω's, for $k = 1, \cdots, 5$, where $\omega_5 = \omega_0$. As a consequence of Lemma 1, we note that $h(w_k) = \omega_k$, $k = 0, \cdots, 4$.

Let $g: D \to \mathbb{C}$ and $\tilde{g}: B \to \mathbb{C}$ be the stereographic projections (from the north pole) of the Gauss map of S_0 when S_0 is parametrized by X and $\tilde{X}$, respectively (see Fig. 3). If we write the components of $\tilde{X}$ as $\tilde{x}$, $\tilde{y}$, $\tilde{z}$, then

$$g(w) = z_w(w)/(x_w(w) - iy_w(w)) \quad \text{for } w \in D,$$

$$\tilde{g}(\omega) = \tilde{z}_\omega(\omega)/(\tilde{x}_\omega(\omega) - i\tilde{y}_\omega(\omega)) \quad \text{for } \omega \in B.$$

Further, g and $\tilde{g}$ are meromorphic in D and B, respectively. For the parametrizations

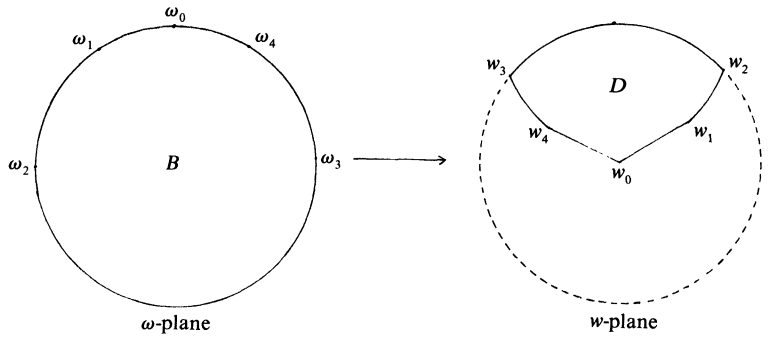


FIG. 3

we have chosen, the Gauss map of each point of S_0 lies in the lower hemisphere, and so g and $\tilde{g}$ are analytic maps from D and B, respectively, into B. Using continuation by reflection and [B], we see that g and $\tilde{g}$ may be extended to be the stereographic projections of the Gauss map of S and that $g \in C^0(\bar{D})$, $\tilde{g} \in C^0(\bar{B})$, and $g(w) = \tilde{g}(h(w))$ for $w \in D$.

LEMMA 1. *$\tilde{g}(\omega) = h^{-1}(\omega)$ for $\omega \in \bar{B}$ and $g(w) = w$ for $w \in \bar{D}$.*

Proof. As noted earlier, the unit normals to Γ_k lie in the plane H_k, $k = 1, 2, 4, 5$, and so $\tilde{g}$ maps s_k into T_k, $k = 1, \cdots, 5$. Using [B], we see that $\tilde{g}$ maps ω_k into ω_k, $k = 0, \cdots, 4$. If we consider various planes and apply the maximum principle (cf. [EL]), we see that $\tilde{g}$ maps s_k onto σ_k, $k = 1, 2, 4, 5$. (These planes are the xy-plane, the plane through Γ_1 and Γ_2, and the plane through Γ_3 and Γ_4.) From [L1], we see that $\tilde{g}(s_3) = \sigma_3$ and $\tilde{g}$ is injective on s_3. Thus $\tilde{g}$ maps ∂B onto ∂D.

Note that the winding number of $\tilde{g}(\partial B)$ about each point of D is 1. By the argument principle, $\tilde{g}$ takes on each value in D exactly once and does not take any values outside D; thus $\tilde{g}$ is a conformal map of B onto D. By the Osgood-Carathéodory Theorem [H, p. 346], $\tilde{g}$ is a homeomorphism of $\bar{B}$ onto $\bar{D}$. Now $\tilde{g}(\omega_k) = w_k$, $k = 0, \cdots, 4$, and $h(w_k) = \omega_k$ for $k = 0, 2, 3$, so $\tilde{g} = h^{-1}$. Furthermore, $g(w) = \tilde{g}(h(w)) = w$ for $w \in \bar{D}$. □

We are interested in the Weierstrass f and g representations of S. Define

$$f(w) = x_w(w) - iy_w(w) \quad \text{for } w \in D,$$

$$\tilde{f}(\omega) = \tilde{x}_\omega(\omega) - i\tilde{y}_\omega(\omega) \quad \text{for } \omega \in B.$$

Then f and $\tilde{f}$ are analytic on D and B, respectively, and the Weierstrass f and g representations of X and of $\tilde{X}$, respectively, are (f, g) and $(\tilde{f}, \tilde{g})$. Note $f(w) = \tilde{f}(h(w))h'(w)$ for $w \in D$. Since X and $\tilde{X}$ can be extended by harmonic continuation, we see that $f \in A(\bar{D}\backslash\{w_0, \cdots, w_4\})$.

In the next two lemmas, we obtain information about f, $\tilde{f}$ that is required for a precise formulation of a Riemann-Hilbert problem for f in the next section.

LEMMA 2. *$\tilde{f}$ has a simple zero at $-i$ and f has a simple zero at i.*

Proof. We may extend Y (from Proposition 1) by reflection across $\partial'' B^+$. Since $Y_\omega(0) = 0$, $h_1(i) = 0 = h_2(-i)$, $X_w(w) = Y_\omega(h_1(w))h_1'(w)$ for $w \in \bar{D}\backslash\{w_0, \ldots, w_4\}$, and $\tilde{X}_\omega(\omega) = Y_\omega(h_2(\omega))h_2'(\omega)$ for $\omega \in \bar{B}\backslash\{\omega_0, \cdots, \omega_4\}$, we see that $X_w(i) = 0$ and $\tilde{X}_\omega(-i) = 0$ and so $f(i) = 0$ and $f(-i) = 0$.

Let us write $Y(\omega) = (x^*(\omega), y^*(\omega), z^*(\omega))$. From [L2], we know that

$$(z^* + ix^*)(\omega) = (H(\omega))^2,$$

where $H(0)=0$ and

$$DH(0)=\operatorname{diag}(e,e), \qquad e\in\Re\backslash\{0\}.$$

If we write $H(\omega)=H_1(u,v)+iH_2(u,v)$, then $z^*(\omega)=(H_1(u,v))^2-(H_2(u,v))^2$ and

$$\begin{aligned}\operatorname{Re}(z^*_{\omega\omega}(\omega))&=\tfrac{1}{4}(z^*_{uu}-z^*_{vv})(\omega)\\&=\tfrac{1}{2}\{H_1H_{1uu}+(H_{1u})^2-H_2H_{2uu}-(H_{2u})^2\\&\qquad -H_1H_{1vv}-(H_{1v})^2+H_2H_{2vv}+(H_{2v})^2\}(\omega).\end{aligned}$$

As $\omega\in B^+$ approaches 0, $\operatorname{Re}(z^*_{\omega\omega}(\omega))$ approaches $\frac{1}{2}\{(H_{1u}(0))^2-(H_{2u}(0))^2-(H_{1v}(0))^2+(H_{2v}(0))^2\}=e^2$ (since in [G, p. 279], it is shown that the second partials of H_1, H_2 are $o(|\omega|^{-1})$) and so $z^*_{\omega\omega}(0)\neq 0$.

Now $z(w)=z^*(h_1(w))$ and so $z_{ww}(w)=z^*_{ww}(h_1(w))(h_1'(w))^2+z^*_w(h_1(w))h_1''(w)$. Then $z_{ww}(i)\neq 0$, and similarly $\tilde{z}_{ww}(-i)\neq 0$. Since $f(w)g(w)=z_w(w)$, $f_w(w)g(w)+f(w)g_w(w)=z_{ww}(w)$ and so $f_w(i)=z_{ww}(i)\neq 0$. Similarly, $\tilde{f}_{\omega}(-i)\neq 0$. □

Let $\lambda_k\pi$ be the angle formed by the intersection of Γ_k and Γ_{k+1}, $k=0,\cdots,4$, where $\Gamma_0=\Gamma_5$ and $\Gamma_3=\{(0,0,z)|z\leqq\phi(A_2)\}$. Let $\delta_k\pi$ be the angle at w_k formed by the tangent lines to σ_k and σ_{k+1} that D "fills."

LEMMA 3. *$\tilde{f}$ has an integrable singularity at ω_k for each $k\in\{0,\cdots,4\}$. In fact, $\tilde{f}(\omega)=O(|\omega-\omega_k|^{\lambda_k-1})$ and $(\omega-\omega_k)^{1-\lambda_k}\tilde{f}(\omega)\to l_k\in\mathbb{C}\backslash\{0\}$ as $\omega\in\bar{B}\backslash\{\omega_k\}$ approaches ω_k, for each $k=0,\cdots,4$.*

Proof. Let $T:\Re^3\to\Re^3$ be the rigid motion sending $\tilde{X}(\omega_k)$ to the origin that maps the line segments forming the angle $\lambda_k\pi$ of Γ_k and Γ_{k+1} at $\tilde{X}(\omega_k)$ into the upper half of the plane $z=0$ as follows: (a) the images of the line segments form equal angles with the y-axis and (b) $T\circ\tilde{X}$ maps s_{k+1} into the first quadrant of the xy-plane. Let us write $T(\xi)=L_k(\xi+r_k)$ for $\xi\in\Re^3$, where $r_k\in\Re^3$ is fixed and L_k is a rotation of $\Re^3$. For $k=0$, $r_0=(-A_0,0)$, and, in matrix form, $L_0=\operatorname{diag}(-1,1,-1)$.

Let h_k be the conformal map of B^+ onto $B(0,1)$ with $h_k(0)=\omega_k$ and $h_k'(0)>0$. Let us now fix $k\in\{0,\cdots,4\}$ and set $Z=T\circ\tilde{X}\circ h_k$. Then $Z(\omega)=(\hat{x}(\omega),\hat{y}(\omega),\hat{z}(\omega))$ maps B^+ onto $T(S)$ and sends $0+0i$ to $(0,0,0)$. Set $\lambda=\lambda_k$. By Theorem 3 of [D],

$$\begin{aligned}\hat{x}_{\omega}(\omega)&=\omega^{\lambda-1}H_2(\omega)+\omega^{-\lambda}H_3(\omega),\\\hat{y}_{\omega}(\omega)&=-i\omega^{\lambda-1}H_2(\omega)+i\omega^{-\lambda}H_3(\omega),\\\hat{z}_{\omega}(\omega)&=H_1(\omega)\end{aligned}$$

for $\omega\in\bar{B}^+\backslash\{0\}$ in the neighborhood of zero,

$$\omega(H_1(\omega))^2+4H_2(\omega)H_3(\omega)\equiv 0,$$

$H_2(0)\neq 0$, and $H_3(0)=0$. If we define $F(\omega)=(\tilde{x}\circ h_k)_{\omega}(\omega)-i(\tilde{y}\circ h_k)_{\omega}(\omega)$, then $F(\omega)=c\omega^{\lambda-1}+o(1)$ for some constant $c\neq 0$. When $k=0$, we remark that $F(\omega)=\omega^{\lambda-1}H_2(\omega)$. Since h_k maps an analytic portion of B^+ into an analytic portion of B, we see that $\tilde{f}(\omega)=0(|\omega-\omega_k|^{\lambda_k-1})$ and $\tilde{f}(\omega)=l_k(\omega-\omega_k)^{\lambda_k-1}+O(|\omega-\omega_k|^{\lambda_k})$ for a constant $l_k\neq 0$. □

Now h is the conformal map of D onto B that maps w_k to ω_k, $k=0,\cdots,4$. We see that for $w\in\bar{D}\backslash\{w_k\}$ near w_k, $\delta=\delta_k$, $\varepsilon=1/\delta_k$,

$$\begin{aligned}h(w)-\omega_k&=c_k(w-w_k)^{\varepsilon}+o(|w-w_k|^{\varepsilon}),\\h'(w)&=\frac{c_k}{\delta}(w-w_k)^{\varepsilon-1}+o(|w-w_k|^{\varepsilon-1})\end{aligned}$$

for some $c_k \in \mathbb{C}\backslash\{0\}$ (e.g., [H, p. 359]). Furthermore,

$$(h^{-1})(\omega) - w_k = e_k(\omega - \omega_k)^{\sigma} + O(|\omega - \omega_k|^{\sigma}),$$

$$(h^{-1})'(\omega) = e_k\delta(\omega - \omega_k)^{\delta-1} + o(|\omega - \omega_k|^{\delta-1})$$

for $\omega \in \bar{B}\backslash\{\omega_k\}$ near ω_k, where $e_k \in \mathbb{C}\backslash\{0\}$. Since $f(w) = \tilde{f}(h(w))h'(w)$, Lemma 3 gives, for $v = -1 + \lambda_k/\delta$,

$$f(w) = c_k(w - w_k)^{v} + o(|w - w_k|^{v})$$

as $w \in \bar{D}\backslash\{w_k\}$ approaches w_k and $f(h^{-1}(\omega)) = c_k(\omega - \omega_k)^{v\delta} + o(|\omega - \omega_k|^{v\delta})$ as $\omega \in \bar{B}\backslash\{\omega_k\}$ approaches ω_k for some $c_k \in \mathbb{C}\backslash\{0\}$.

Note that $\lambda_0\pi = 2\alpha$ and $\delta_0\pi = \pi - 2\alpha$; $\lambda_1\pi = \cos^{-1}(\cos(\gamma)\cos(\beta))$ and $\delta_1\pi > \pi/2$; $\lambda_2\pi = \pi/2 - \beta$ and $\delta_2\pi > \pi/2$; $\lambda_3\pi = \pi/2 - \beta$ and $\delta_3 > \pi/2$; and $\lambda_4\pi = \cos^{-1}(\cos(\gamma)\cos(\beta))$ and $\delta_4\pi > \pi/2$. Thus $f(h^{-1}(\omega))$ has an integrable singularity at ω_1, ω_2, ω_3, and ω_4. Also, at ω_0, $f(h^{-1}(\omega)$ has an integrable singularity if $\alpha < \pi/4$ and has a "fractional zero" if $\alpha > \pi/4$.

2. The Hilbert problem. We wish to derive a homogeneous boundary value problem for which $f(w)$ is a solution. Let $f(u+iv) = f_1(u, v) + if_2(u, v)$, where f_1 and f_2 are real-valued. Since $g(w) = w$,

$$x_w(w) = f(w)(1 - w^2)/2,$$

$$y_w(w) = if(w)(1 + w^2)/2,$$

$$z_w(w) = wf(w)$$

for $w \in D$. Since $d/dw = \frac{1}{2}((\partial/\partial u) - i(\partial/\partial v))$, taking real and imaginary parts gives us

$$x_u(u, v) = (1 - u^2 + v^2)f_1 + 2uvf_2,$$

$$x_v(u, v) = 2uvf_1 - (1 - u^2 + v^2)f_2,$$

$$y_u(u, v) = -2uvf_1 + (1 + u^2 - v^2)f_2,$$

$$y_v(u, v) = -(1 + u^2 - v^2)f_1 + 2uvf_2,$$

$$z_u(u, v) = 2(uf_1 - vf_2),$$

$$z_v(u, v) = 2(-vf_1 - uf_2).$$

Let us set $m_1 = \cot(\alpha)$, $m_2 = \cot(\alpha + \gamma)$, $m_4 = -\cot(\alpha + \gamma)$, $m_5 = -\cot(\alpha)$, $n_1 = 0$, $n_2 = -\tan(\beta)\csc(\alpha + \gamma)$, $n_4 = \tan(\beta)\csc(\alpha + \gamma)$, and $n_5 = 0$. The condition $(x, y, z) \in \Gamma_k$ is equivalent to $y = m_k x + b_k$, $z = n_k x + c_k$, $k = 1, 2, 4, 5$, where (b_k, c_k) are constants depending on α, β, γ, k. We know that X maps σ_k into Γ_k, $k = 1, 2, 4, 5$, and maps σ_3 into the z-axis. If we parametrize σ_k by $w_k(t) = (u_k(t), v_k(t))$, then $X(w_k(t)) \in \Gamma_k$ and so $y(w_k(t)) = m_k x(w_k(t)) + b_k$, $z(w_k(t)) = n_k x(w_k(t)) + c_k$ for $k = 1, 2, 4, 5$ and $x(w_3(t)) = y(w_3(t)) = 0$. Differentiating with respect to t yields the two equivalent equations $y_u u_k' + y_v v_k' = m_k(x_u u_k' + x_v v_k')$, $z_u u_k' + z_v v_k' = n_k(x_u u_k' + x_v v_k')$ when $k = 1, 2, 4, 5$ and $x_u u_3' + x_v v_3' + y_v v_3' = 0$ when $k = 3$. If we now write $x_u, \cdots, z_v$ in terms of f_1, f_2, u, v and rearrange each equation slightly, we obtain

$$a_k(u, v)f_1(u, v) + b_k(u, v)f_2(u, v) = 0$$

when $(u, v) \in \sigma_k$, for $k = 1, \cdots, 5$, where

$$a_1(u, v) = -\cos(2\alpha),$$

$$a_2(u, v) = 2uv(v - v_0) - (1 + u^2 - v^2)(u - u_0) - \cot(\alpha + \gamma)(1 - u^2 + v^2)(v - v_0) + 2\cot(\alpha + \gamma)uv(u - u_0),$$

$$a_3(u, v) = 2v(u^2 - v^2),$$

$$a_4(u, v) = 2uv(v - v_0) - (1 + u^2 - v^2)(u - u_0) + \cot(\alpha + \gamma)(1 - u^2 + v^2)(v - v_0) - 2\cot(\alpha + \gamma)uv(u - u_0),$$

$$a_5(u, v) = \cos(2\alpha),$$

$$b_1(u, v) = \sin(2\alpha),$$

$$b_2(u, v) = (1 + u^2 - v^2)(v - v_0) + 2uv(u - u_0) - 2\cot(\alpha + \gamma)uv(v - v_0) - \cot(\alpha + \gamma)(1 - u^2 + v^2)(u - u_0),$$

$$b_3(u, v) = -4uv^2,$$

$$b_4(u, v) = (1 + u^2 - v^2)(v - v_0) + 2uv(u - u_0) + 2\cot(\alpha + \gamma)uv(v - v_0) + \cot(\alpha + \gamma)(1 - u^2 + v^2)(u - u_0),$$

$$b_5(u, v) = \sin(2\alpha).$$

We point out that if $Z : \bar{D} \to R^3$ is C^1 on $\bar{D}$ except at a finite number of points of D, $Z \in C^0(\bar{D})$, and $Z(w) = (p(w)(1 - w^2)/2,\ ip(w)(1 + w^2)/2,\ wp(w))$, where $p(w) = p_1(u, v) + ip_2(u, v)$, then the condition $a_k p_1 + b_k p_2 = 0$ on σ_k is equivalent to $Z(\sigma_k)$ being a line segment parallel to Γ_k if $k \neq 3$ and parallel to the z-axis if $k = 3$. The forthcoming modification of a_2, a_4, b_2, b_4 will not change this observation.

In the theory of Hilbert problems with discontinuous coefficients, the coefficient $G = a - ib$ must be bounded away from 0 [M, pp. 42–53]. In our case $a_2(w_2) = b_2(w_2) = 0$ and $a_4(w_3) = b_4(w_3) = 0$ and so some modification is required. If $(u, v) \in \sigma_2$, then $u = -u_0 + r\cos(\theta)$ and $v = v_0 + r\sin(\theta)$ for some $\theta \in (-\pi/2, \pi/2)$. Set $t = \tan(\theta/2)$ and note that $\cos(\theta) = (1 - t^2)/(1 + t^2)$, $\sin(\theta) = 2t/(1 + t^2)$, $t \in [t_0, t_1]$, where $t_0 = \tan(\theta_0/2)$, $t_1 = \tan(\theta_1/2)$, and θ_0, θ_1 are in $(-\pi/2, \pi/2)$. Let $c(t) = a_2(w_2(t))$, $d(t) = b_2(w_2(t))$, and $e(t)$ be the greatest common divisor of $c(t)$ and $d(t)$. Let us define new functions $a_2(u, v)$ and $b_2(u, v)$ for $(u, v) \in \sigma_2$ by setting $a_2(w_2(t)) = c(t)/e(t)$ and $b_2(w_2(t)) = d(t)/e(t)$. We may repeat this process and replace a_4 and b_4.

Define $a(u, v)$ and $b(u, v)$ on $\partial D \backslash \{w_0, \cdots, w_4\}$ by setting $a = a_k$ and $b = b_k$ on $\sigma_k \backslash \{w_{k-1}\}$. Note that each a_k and b_k is continuous on σ_k and so a and b have one-sided limits at $w_0, \cdots, w_4$. Also a and b are certainly piecewise Lipschitz continuous on ∂D. Now a and b are discontinuous at $w_1, \cdots, w_4$ and at w_0 unless $\alpha \geqq \pi/4$. Set $G(w) = a(w) - ib(w)$.

Let R be defined on ∂D (or almost everywhere on ∂D). Consider the Hilbert problem $H(\cdot, R)$ of finding a function analytic in D that satisfies $\operatorname{Re}(G\psi) = R$ on $\partial D \backslash \{w_0, \cdots, w_4\}$ and that has the "singularity pattern of f" at $w_0, \cdots, w_4$. The "index" of this problem (with indicated behavior at w_k, $k = 0, \cdots, 4$) is 1 and the general solution can be found using [M, pp. 42–53]. If ψ_0 denotes a solution of $H(\cdot, 0)$, then $\psi(w) = \psi_0(w)(c + H(w))$, where H is the analytic function in D whose imaginary part on ∂D is $R/(iG\psi_0)$. This is true since $\operatorname{Re}(G\psi) = c\operatorname{Re}(G\psi_0) + \operatorname{Re}(G\psi_0 H) = 0 + \operatorname{Im}(H) = R$ on ∂D (recall that $\operatorname{Re}(G\psi_0) = 0$ on ∂D). The analytic function H can be found by using the Schwarz formula; formally,

$$H^*(\omega) = \frac{1}{2\pi i}\int_{|t|=1} \frac{R^*(t)(t + \omega)\, dt}{G^*(t)\psi_0^*(t)t(t - \omega)}$$

for $\omega \in \partial B$, where $H^*(\omega) = H(h^{-1}(\omega))$ and similarly for R^*, G^*, ψ^*. For later use, set $\alpha_k = \lambda_k - \delta_k$ for $k = 0, \cdots, 4$ and note $\alpha_k < 0$ if $k = 1, \cdots, 4$.

3. Bernstein functions. We will now formulate a boundary value problem that will lead to the surfaces to be constructed. Let $A_3^* = (x_3^*, y_3^*)$ be a point on $\partial\Omega$ strictly between A_3 and A_4. Set $A_2^* = (x_2^*, y_2^*) = (-x_3^*, y_3^*) \in \partial\Omega$. Let K be a C^2 curve in $\mathfrak{R}^2$ between A_2^* and A_3^* such that K is symmetric with respect to the y-axis, K and $\partial\Omega$ are tangent at A_2^* and A_3^*, and K is strictly concave with respect to $\{(x, y+t) \mid (x, y) \in K, t \leqq 0\}$. Let us parametrize K by $(x, k(x))$, $x_2^* \leqq x \leqq x_3^*$. We know that $k''(x) > 0$ unless $x \in E = \{x \in (x_2^*, x_3^*) \mid k''(x) = 0\}$, which we assume is finite, and possibly at x_2^* and x_3^*. We require $0 \notin E$. We will impose additional restrictions on k near $E \cup \{x_2^*, x_3^*\}$.

We wish to parametrize K in a special way. Since K is strictly concave and is tangent to $\partial\Omega$ at A_2^* and A_3^*, we see that for each $e^{i\theta} \in \sigma_3$ there is a unique point $(A(\theta), B(\theta))$ on K such that the tangent line to K at $(A(\theta), B(\theta))$ is orthogonal to $(\cos(\theta), \sin(\theta))$, and conversely. The parametrization $(A(\theta), B(\theta))$, $\alpha + \gamma \leqq \theta \leqq \pi - (\alpha + \gamma)$, is continuous, is C^1 on $N = \{\theta \in (\alpha + \gamma, \pi - (\alpha + \gamma)) \mid A(\theta) \notin E\}$, and satisfies

$$(A'(\theta), B'(\theta)) \cdot (\cos(\theta), \sin(\theta)) = 0$$

and so $B'(\theta) = -\cot(\theta) A'(\theta)$, for $\theta \in N$. The symmetry of K implies $A'(\pi - \theta) = A'(\theta)$ and $B'(\pi - \theta) = -B'(\theta)$. We can easily check that $k'(A(\theta)) = -\cot(\theta)$, which could be taken as a definition of $A(\theta)$, and $A'(\theta) = \csc^2(\theta)/k''(A(\theta))$.

Let us assume that there exists $\beta \in (0, 1)$ and $C_1 > 0$ such that $k''(x) \geqq C_1 |x - \bar{x}|^\beta$, whenever x is sufficiently close to $\bar{x}$, for each $\bar{x} \in E$. If $k''(x_2^*) = k''(x_3^*) = 0$, let us assume that there exist $\gamma \in (0, -\alpha_3/\delta_3)$ and $C_2 > 0$ such that $k''(x) \geqq C_2 |x - x_3^*|^\gamma$ whenever $x \leqq x_3^*$ is sufficiently near x_3. Let us examine $A(\theta)$ and $A'(\theta)$. If $k''(A(\theta^*)) = 0$, then $A'(\theta) \to \infty$ as $\theta \to \theta^*$ and so $|A(\theta) - A(\theta^*)| \geqq 2|\theta - \theta^*|$ for $\theta \in (\theta_2^*, \theta_3^*)$ near θ^*. Now if $w = e^{i\theta}$ and $w^* = e^{i\theta^*}$ then

$$|A(\theta) - A(\theta^*)| \geqq 2|\arg(w) - \arg(w^*)| = 2|\mathrm{Ln}(w) - \mathrm{Ln}(w^*)| \geqq |w - w^*|$$

and $A'(\theta) = \csc^2(\theta)/k''(A(\theta)) \leqq 2\csc^2(\theta^*) C_1 |A(\theta) - A(\theta^*)|^{-\beta} \leqq C' |w - w^*|^{-\beta}$ if w is near w^*, and similarly for w_2 and w_3.

Finally let us assume that $k \in C^{2,\delta}((x_2, x_3))$ for some $0 < \delta \leqq \min\{\beta, \gamma\}$. Then $|A'(\lambda_2) - A'(\lambda_1)| \leqq C(\lambda_1, \lambda_2) |\lambda_2 - \lambda_1|^\delta$ for all $\lambda_1, \lambda_2 \in (\theta_2^*, \theta_3^*)$ with $\lambda_1, \lambda_2 \notin N$, and for λ_1, λ_2 in a compact subset of N, $C(\lambda_1, \lambda_2)$ is uniformly bounded.

Define $R(w) = 0$ if $w \notin \sigma_3$ and $R(w) = A'(\theta)$ if $e^{i\theta} = w \in \sigma_3$ and $k''(A(\theta)) \neq 0$. Let $R^*(\omega) = R(h^{-1}(\omega))$, $G^*(\omega) = G(h^{-1}(\omega))$, $\psi_0 = f$, and $\psi_0^* = \tilde{f}$. Recall that $\tilde{f}(\omega) = (\omega + i) e(\omega)$ with $e(-i) \neq 0$ and, for $k = 2, 3$,

$$\tilde{f}(\omega) = c_k (\omega - \omega_k)^{\alpha_k} + O(|\omega - \omega_k|^{\alpha_k}).$$

Note that $R^*(\omega)$ is Hölder continuous at ω_2 and ω_3 with exponent $-\alpha_3 - (\gamma/\delta_3) > 0$ (use [H, p. 359] and $R^*(\omega) = O(|\omega - \omega^*|^{-\beta})$ if $\omega^* \in E^* = \{\omega \in s_3 \mid \arg(h^{-1}(\omega)) \notin N\}$). Finally R^* is Hölder continuous on $\partial B \backslash E \cup \{-i\}$ and uniformly Hölder continuous on each compact subset of $\partial B \backslash E \cup \{-i\}$.

Define

$$H^*(\omega) = \frac{1}{2\pi i(\omega + i)} \left[\text{P.V.} \int_{\partial B} \frac{R^*(t)(t+\omega)\, dt}{G^*(t) e(t) t(t-\omega)} - \text{P.V.} \int_{\partial B} \frac{R^*(t)(t+\omega)\, dt}{G^*(t) e(t) t(t+i)} \right]$$

for all $\omega \in \bar{B} \backslash E^*$. Since

$$\frac{\omega + i}{(t+i)(t-\omega)} = \frac{1}{t-\omega} - \frac{1}{t+i},$$

we see that

$$H^*(\omega)=\frac{1}{2\pi i}\,\text{P.V.}\int_{\partial B}\frac{R^*(t)(t+\omega)\,dt}{G^*(t)f(t)t(t-\omega)(t+i)}$$

for $\omega\in\bar{B}\backslash E^*\cup\{-i\}$, where we take principal values at $-i$ and at ω if $\omega\in\partial B$. Using Theorem 14.1c of [H], we see that H^* is analytic in B and continuous on $\bar{B}\backslash E^*\cup\{-i\}$. Furthermore,

$$\text{Im}\,(H^*(\omega))=R^*(\omega)/(iG^*(\omega)\tilde{f}(\omega))$$

for each $\omega\in\partial B\backslash E^*\cup\{-i\}$ [H, pp. 100-103] and $H(\omega)=O(|\omega-\omega^*|^{-\beta})$ for $\omega\in\bar{B}$ near $\omega^*\in E^*$, as a simple argument shows.

Since $H(w)=H^*(h(w))$ is continuous on $\bar{D}\backslash\{e^{i\theta}\in\sigma_3\,|\,A(\theta)\in E\}\cup\{i\}$, we can define

$$c_0=c_0(K,D)=-\inf\{\text{Re}\,(H(w))\,|\,w\in\sigma\backslash\sigma_3\}.$$

Since $R=0$ on $\sigma\backslash\sigma_3$, $\text{Im}\,(H(w))=0$ if $w\in\sigma\backslash\sigma_3$ and so

$$c_0=-\inf_{\sigma\backslash\sigma_3}H=\sup_{\sigma\backslash\sigma_3}(-H).$$

Note that if $c\geqq c_0$, $c+H(w)>0$ for w at all but a finite number of points of $\sigma\backslash\sigma_3$.

DEFINITION. For a curve K as above and any $\ell>\sqrt{(x_2^*)^2+(y_2^*)^2}$, define $\Omega(K,\ell)$ as the domain bounded by K and the line segments between A_2^* and ℓA_1, ℓA_1 and ℓA_0, ℓA_0 and ℓA_4, and ℓA_4 and A_3^* (see Fig. 4).

Set $\phi_\ell(x,y)=\ell\phi(x/\ell,y/\ell)$ for $(x,y)\in\partial\Omega(k,\ell)$.

PROBLEM $B(K,\ell)$. For a curve K as above and $\ell>\sqrt{(x_2^*)^2+(y_2^*)^2}$, find $F^+\in C^2(\Omega(K,\ell))\cap C^0(\overline{\Omega(K,\ell)})$ that satisfies the minimal surface equation in $\Omega(K,\ell)$ with $F^+=\phi_\ell$ on $\partial\Omega(K,\ell)\backslash K$ such that as $(x,y)\in\Omega(K,\ell)$ approaches $(x_0,y_0)\in\{(x,y)\in K\,|\,x\notin E\}$, the exterior normal derivative of F^+ at (x,y) approaches ∞.

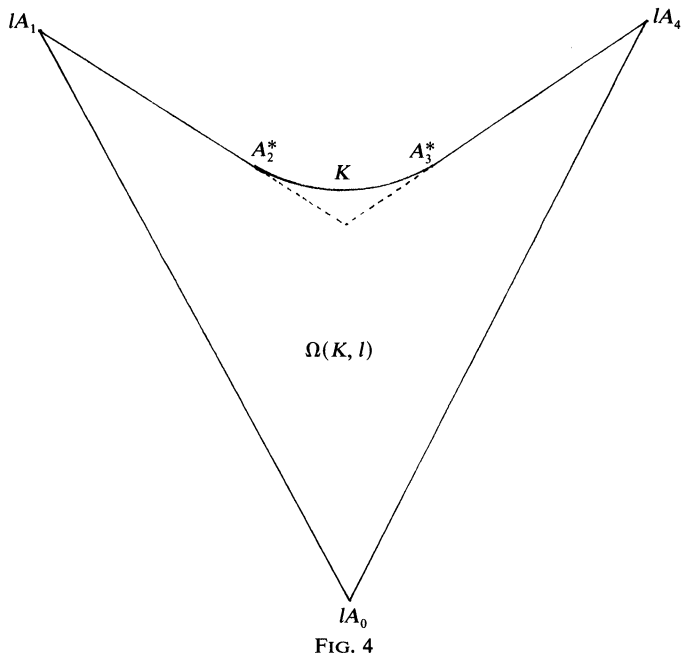


FIG. 4

We will find a number $\ell_0 = \ell_0(K)$ such that $B(K, \ell)$ has a (necessarily unique) solution if $\ell \geqq \ell_0$. If $\ell < \ell_0$, we obtain a candidate that does not solve $B(K, \ell)$.

For c real, define $\psi(w, c) = f(w)(c + H(w))$ and note that $\psi(\cdot, c)$ solves $H(\cdot, R)$ (on $h^{-1}(E^*)$, R is not defined). Clearly, f is continuous at i. Define

$$x(w, c) = \operatorname{Re}\left(\int_0^w \frac{1}{2}\psi(t, c)(1 - t^2)\, dt\right)$$

for $w \in \bar{D}$. Set $\ell_0 = \ell_0(K) = \frac{1}{2}\csc(\alpha + \gamma)(x(w_4, c_0) - x(w_1, c_0))$ and define $c(\ell) = c(K, \ell) = \ell - \ell_0(K) + C_0(K)$. Define

$$y(w, c) = \operatorname{Re}\left(\int_0^w \frac{i}{2}\psi(t, c)(1 + t^2)\, dt\right) - (c - c_0(K) + \ell_0(K)) \sin(\alpha)/\sin(\gamma),$$

$$z(w, c) = \operatorname{Re}\left(\int_0^w t\psi(t, c)\, dt\right),$$

$$X(w, c) = (x(w, c), y(x, c), z(w, c))$$

for $w \in \bar{D}$. Note that $X(\cdot, c) \in C^0(\bar{D})$ and $X(w, c) = X(w, c_0) + (c - c_0)X(w)$. In particular, $x(w_4, c(\ell)) - x(w_1, c(\ell)) = x(w_4, c_0) - x(w_1, c_0) + (c(\ell) - c_0)(x(w_4) - x(w_1)) = 2\ell_0 \sin(\alpha + \gamma) + 2(c(\ell) - c_0)\sin(\alpha + \gamma) = 2\ell \sin(\alpha + \gamma)$.

Note that

$$X_w(w, c) = (c + H(w))X_w(w)$$

for $w \in \bar{D} \backslash h^{-1}(E^*)$ and $c + H(w)$ is real and positive if $w \in \sigma \backslash \sigma_3$ and $c > c_0$ ($c = c_0$ implies $c + H(w) > 0$ on $\sigma \backslash \sigma_3$ except at a finite number of points). Thus the sign pattern of $x_u(\cdot, c), \cdots, z_v(\cdot, c)$ is the same as that of $x_u, \cdots, z_v$ on $\sigma \backslash \sigma_3$ if $c \geqq c_0$.

Suppose $w = e^{i\theta} \in \sigma_3 \backslash (h^{-1}(E^*) \cup \{w_2, w_3\})$. Since

$$a(w)f_1(w, c) + b(w)f_2(w, c) = A'(\theta),$$

$$a(w)f_1(w, c) + b(w)f_2(w, c) = \frac{d}{d\theta}(x(\cos(\theta), \sin(\theta), c)),$$

where $f = f_1 + if_2$ and $c \in R$, we see that $d/d\theta\,(x(\cos(\theta), \sin(\theta), c)) = A'(\theta)$.

Since $x(\cdot, \cdot, c)$ is continuous, we obtain $x(\cos(\theta), \sin(\theta), c) = A(\theta) + A_0$, for $\alpha + \gamma \leqq \theta \leqq \pi - (\alpha + \gamma)$. One unit normal to $X(\bar{D}, c)$ at $w = e^{i\theta}$ with $\theta \in N$ is $(\cos(\theta), \sin(\theta), 0)$. Since $d/d\theta\,(X(\cos(\theta), \sin(\theta), c))$ lies in the tangent plane of $X(\bar{D}, c)$ at w, we see that $d/d\theta\,(y(\cos(\theta), \sin(\theta), c)) = -\cot(\theta)A'(\theta) = B'(\theta)$. Thus $y(\cos(\theta), \sin(\theta), c) = B(\theta) + B_0$ and the projection onto the xy-plane of $X(\sigma_3, c)$ is the translation of K by (A_0, B_0).

Suppose $w \in \sigma \backslash \sigma_3$. Then $R(w) = 0$ and $a(w)f_1(w, c) + b(w)f_2(w, c) = 0$. As we indicated in § 2, $X(\sigma_k, c)$ lies on a line parallel to Γ_k, $k = 1, 2, 4, 5$. If $\ell \geqq \ell_0(K)$, then $c(\ell) \geqq c_0$ and so $X(\cdot, c(\ell))$ maps ∂D strictly monotonically onto the Jordan curve $X(\partial D, c(\ell))$. Now $H_2(u, v) = \operatorname{Im}(H^*(h(u + iv)))$ is odd on ∂D so $f_1(u, v, c(\ell))$ is an even function of u and $f_2(u, v, c(\ell))$ is an odd function of u for $(u, v) \in \bar{D}$. Thus $x(u, v, c(\ell))$ is odd in u, $y(u, v, c(\ell))$ and $z(u, v, c(\ell))$ are even in u (and $X(\partial D, c(\ell)$ is a simple Jordan curve that is symmetric with respect to the yz-plane and has a simple projection on the xy-plane). Also $X(0, c(\ell)) = \ell(A_0, 0)$ and $x(w_4, c(\ell)) - x(w_1, c(\ell)) = 2\ell \sin(\alpha + \gamma)$ implies $X(\sigma_1, c(\ell)) = \ell\Gamma_1$, $X(\sigma_5, c(\ell)) = \ell\Gamma_5$, and the projection of $X(\partial D, c(\ell))$ on the xy-plane is $\partial\Omega(K, \ell)$. Using [R, p. 36], we see that $X(D, c(\ell))$ has no interior branchpoints and, since $g(w) = w$, $X(D, c(\ell))$ must be the graph of a function F^+. From our discussion, we see that F^+ is a solution of $B(K, \ell)$.

We give the following example to illustrate the applicability range of these theorems. We compare the results obtained using our methods with those obtained using catenoids, which are standard comparison surfaces.

Example. Suppose $\alpha = \pi/8$, $\gamma = \pi/8$, and $\beta = \pi/4$. Note that $A_0 = (0, -1)$, $A_1 = (-1/\sqrt{2}, 1/\sqrt{2})$, and $A_4 = (-1/\sqrt{2}, 1/\sqrt{2})$. Let $\varepsilon \in (0, \beta/2\pi]$; say, $\varepsilon = 0.1$. Define $k \in C^{2,0.1}([-\frac{1}{2}, \frac{1}{2}])$ by $k''(x) = c(1-4x^2)^{\varepsilon}$, $k'(0) = 0$, and $k(-\frac{1}{2}) = k(\frac{1}{2}) = \frac{1}{2}$, where $1/c = \int_0^{0.5} (1-4t^2)^{\varepsilon}\, dt$. Note that $k'(-\frac{1}{2}) = -1$ and $k'(\frac{1}{2}) = 1$; also, $k(x)$ satisfies the conditions mentioned at the beginning of § 3. Let $K = \{(x, k(x)): -\frac{1}{2} \leqq x \leqq \frac{1}{2}\}$ and set $\ell_0 = \ell_0(K)$. Let $U = \Omega(K, \ell_0)$.

We are interested in two questions. First, what information can we obtain about the existence of a solution of problem $B(K, \ell_0)$ using our construction and catenoids? Second, if $\xi \in C^2(U) \cap C^0(\bar{U} \backslash K)$ is a solution of the minimal surface equation in U and $\xi \leqq \phi_{\ell_0}$ on $\partial U \backslash K$, what information can we obtain concerning the behavior of ξ near $(\pm\frac{1}{2}, \frac{1}{2})$, in particular, concerning the lim sup of $\xi(x, y)$ as $(x, y) \in \bar{U}$ approaches $(\pm\frac{1}{2}, \frac{1}{2})$, using our construction and using catenoids? With respect to our construction, we saw that a solution $F \in C^0(\bar{U})$ of $B(K, \ell_0)$ exists. Also, from § 4 we see that $\xi(x, y) \leqq F(x, y)$ for $(x, y) \in \bar{U} \backslash K$ and so $\limsup_{(x,y)\to(\pm 1/2, 1/2)} \xi(x, y) \leqq F(\pm\frac{1}{2}, \frac{1}{2}) = \ell_0 - 1/\sqrt{2}$. With respect to using catenoids, as the following paragraph indicates, the existence of a bounded solution $\psi \in C^2(U) \cap C^0(\bar{U} \backslash K)$ of the minimal surface equation in U with $\psi = \phi_{\ell_0}$ on $\partial U \backslash K$ and $\partial\psi/\partial\eta = +\infty$ almost everywhere on K cannot be obtained using catenoids; at best, perhaps the existence of $\psi \in BV(\Omega)$ with $\operatorname{tr}(\psi) \in L^1(\partial U)$ might be possible. Furthermore, no information concerning lim sup $\xi(x, y)$ as $(x, y) \in U$ approaches $(\pm\frac{1}{2}, \frac{1}{2})$ can be obtained using this method.

Suppose $x_0 \in (-\frac{1}{2}, \frac{1}{2})$. Let $A(a, b)$ be the annulus with center $(x_0 - ak'(x_0)w(x_0), k(x_0) + aw(x_0))$, inner radius a and outer radius b, where $w(x) = (1 + (k'(x))^2)^{-1/2}$; note that the inner boundary $\partial_a A$ of $A(a, b)$ is tangent to K at $(x_0, k(x_0))$ and the center of $A(a, b)$ lies "above" K. Suppose the outer boundary $\partial_b A$ of $A(a, b)$ does not interest U and $\partial_a A$ lies below K (see Fig. 5). Note then that $a > R(x_0)$ and $b - a > 1/\sqrt{2}$, where $R(x)$ is the radius of curvature of K at $(x, k(x))$ (i.e., $1/R(x) = k''(x)w^3(x)$). Let the graph of $u \in C^2(A(a, b)) \cap C^0(\overline{(A(a, b))})$ be a piece of a catenoid such that $u = 0$ on $\partial_b A$ and $\partial u/\partial n = -\infty$ on $\partial_a A$. Then $\xi(x, y) \leqq u(x, y)$

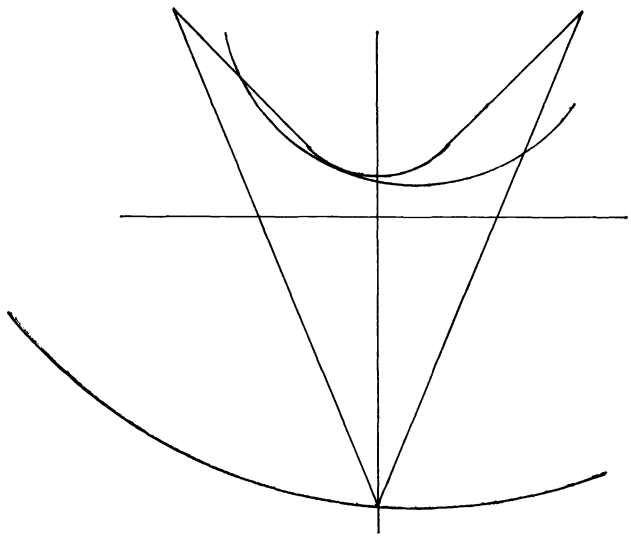


FIG. 5

for $(x, y) \in U \cap A(a, b)$ (e.g., [F1]), provided that $u \geqq \phi_{\ell_0}$ on $\partial U \cap A(a, b)$. Define $\lambda(r; a, b) = a \cosh^{-1}(b/a) - a \cosh^{-1}(r/a)$ for $r \geqq a$ and let $r(x, y)$ be the distance between (x, y) and the center of $A(a, b)$. Then $u(x, y) = \lambda(r(x, y); a, b)$ and $u(x, y) > \lambda(r(x, y); R(x_0), R(x_0) + 1/\sqrt{2})$. Now $u(x, y)$ is (possibly) an upper bound for $\xi(x, y)$ at $(x_0, K(x_0))$ and the least $u(x_0, k(x_0))$ could be is (greater than) $L(x_0)$, where $L(x) = R(x) \cosh^{-1}(1 + 1/\sqrt{2}R(x))$. Thus, the smallest upper bound catenoids could provide for ξ on K is $L(x)$. Since $R(x) \to \infty$ as $x \to \pm\frac{1}{2}$, $L(x) \to \infty$ as $x \to \pm\frac{1}{2}$. In particular, if we take $\xi = F$, we see that the use of catenoids does not imply that F is bounded (much less continuous).

4. Applications. Using the comparison lemma due to Finn [F1] (see also [F2]), we see that the results of the previous section imply the following theorem.

THEOREM. *Let K and $\Omega(K, \ell)$ be as in § 3 with $\ell \geqq \ell_0(K)$. Then there exists $F^+ \in C^0(\overline{\Omega(K, \ell)}) \cap C^2(\Omega(K, \ell))$, a solution of the minimal surface equation in $\Omega(K, \ell)$ with $F^+ = \phi_\ell$ on $\partial\Omega(K, \ell) \setminus K$, such that*

$$f \leqq F^+ \quad \text{on } \Omega(L, \ell)$$

for every solution f of the minimal surface equation in $\Omega(K, \ell)$ with $f \leqq \phi_\ell$ on $\partial\Omega(K, \ell) \setminus K$.

One of the distinctive features of this result is that the curvature of K need not be bounded away from zero; the curvature of K can vanish on E and at the endpoints of K. As we saw in the example in § 3, standard comparison surfaces such as the catenoid are not useful in this case. In fact, the existence theorem we have proven does not follow from known results obtained by the variational theory such as those of [G]. It would be interesting to see if we could apply that theory, using refinements such as those recently reported in [F2], to obtain these results.

Suppose U is a domain in the plane whose boundary consists of a concave, symmetric arc K and a locally convex arc L such that the tangent lines to the endpoints of K make an angle larger than $\pi/2$ and K satisfies the conditions of § 3. Here we may assume that U has been rotated and translated into a suitable position. Then for $\ell \geqq \ell_0(K)$ sufficiently large, $U \subseteq \Omega(K, \ell)$ and $\partial U \cap \partial\Omega(K, \ell) = K$ (see Fig. 6).

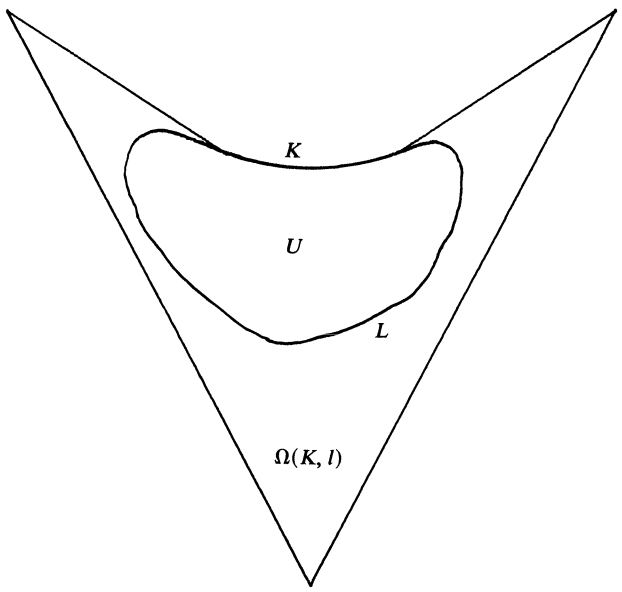


FIG. 6

COROLLARY. *Let U and $\Omega(K, \ell)$ be as above. Let $\psi \in L^\infty(L)$ and let $f \in C^2(U)$ be a solution of the minimal surface equation in U such that $f = \psi$ almost everywhere on L. Then*

$$f \leqq F^+ + \sup_L \psi \quad \text{on } U,$$

where $F^+ = F^+(K, \ell) \in C^0(\overline{\Omega(K, \ell)})$.

The proof of this corollary follows from the general maximum principle and the previous theorem.

Acknowledgment. The authors acknowledge numerous helpful conversations with Bob Gulliver concerning this work and thank him for his criticism and encouragement.

REFERENCES

[B] M. BEESON, *The behavior of a minimal surface in a corner*, Arch. Rational Mech. Anal., 65 (1977), pp. 379-393.

[D] G. DZJUK, *Uber quasilinear elliptische Systeme mit isotherman Parametern on Ecken der Randkurv*, Analysis, 1 (1981), pp. 63-81.

[EL] A. ELCRAT AND K. LANCASTER, *On the behavior of a non-parametric minimal surface in a non-convex quadrilateral*, Arch. Rational Mech. Anal., 94 (1986), pp. 209-226.

[ET] A. ELCRAT AND L. N. TREFETHEN, *Free streamline flow over a polygonal obstacle*, J. Comput. Appl. Math., 14 (1986), pp. 251-265; Numerical Conformal Mapping, L. N. Trefethen, ed., North-Holland, Amsterdam, New York, 1986.

[F1] R. FINN, *Remarks relevant to minimal surfaces and to surfaces of prescribed mean curvature*, J. Analyse Math., 14 (1965), pp. 139-160.

[F2] ———, *Equilibrium Capillary Surfaces*, Springer-Verlag, Berlin, New York, 1986.

[G] E. GIUSTI, *Boundary value problems for non-parametric surfaces of prescribed mean curvature*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 3 (1976), pp. 501-548.

[GL] R. GULLIVER AND F. LESLEY, *On boundary branch points of minimizing surfaces*, Arch. Rational Mech. Anal., 52 (1973), pp. 20-25.

[H] P. HENRICI, *Applied and Computational Complex Analysis*, Vol. 3, John Wiley, New York, 1986.

[L1] K. LANCASTER, *Boundary behavior of a non-parametric minimal surface in $\mathfrak{R}^3$ at a non-convex point*, Analysis, 5 (1985), pp. 61-69.

[L2] ———, *Nonparametric minimal surfaces in $\mathfrak{R}^3$ whose boundaries have a jump discontinuity*, Internat. J. Math. Math. Sci., 11 (1988), pp. 651-656.

[M] V. MONAKHOV, *Boundary-Value Problems with Free Boundaries for Elliptic Systems of Equations*, Amer. Math. Soc. Transl. 57, American Mathematical Society, Providence, RI, 1983.

[N] J. C. C. NITSCHE, *On new results in the theory of minimal surfaces*, Bull. Amer. Math. Soc., 71 (1965), pp. 195-270.

[R] T. RADÓ, *On the Problem of Plateau*, Ergeb. Math. Grenzgeb. (3), Springer-Verlag, Berlin, 1933.

[S] L. SIMON, *Boundary regularity for solutions of the non-parametric least area problem*, Ann. of Math. (2), 103 (1976), pp. 429-455.

[T] L. N. TREFETHEN, *Numerical Conformal Mapping*, North-Holland, Amsterdam, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1069-1080, September 1989

003

APPLICATIONS OF SHILNIKOV'S THEORY TO SEMILINEAR ELLIPTIC EQUATIONS*

C. J. BUDD†

Abstract. This paper shows how techniques from the theory of dynamical systems may be employed to study semilinear elliptic equations having nonlinearities that grow faster than the critical Sobolev exponent. In particular, techniques from Shilnikov's theory are used to study the symmetric solutions of the equation

$$\Delta u + \lambda(u^p + u^q) = 0, \qquad u|_{\partial B} = 0,$$

where B is the unit ball in $\mathbb{R}^3$ and $q < 3 < 5 < p$. In particular, there is shown to be a critical value of λ at which the above equation has an infinite number of positive solutions.

Key words. semilinear elliptic equations, dynamical systems, Shilnikov's theory, bifurcation, critical Sobolev exponents

AMS(MOS) subject classifications. 34B15, 35J25, 35J65

1. Introduction. In this paper we will use methods from the theory of dynamical systems to study the positive solutions of the following differential equation:

$$u_{rr} + \frac{2}{r} u_r + u^p + u^q = 0, \qquad u(0) = N, \quad u_r(0) = 0, \tag{1.1}$$

where $p > 5$ and $q < 3$.

An application of standard theory given, for example, in Smoller and Wasserman [21] or Ni and Serrin [16] shows that $u(r)$ is an analytic function of N and has a first positive zero, defined by $\mu(N) < \infty$, that is also an analytic function of N. We may thus deduce that a solution of problem (1.1) is also a radially symmetric solution of the following partial differential equation problem:

$$\Delta u + u^p + u^q = 0, \qquad u > 0 \quad \text{in } B, \qquad u|\partial B = 0, \tag{1.2}$$

where

$$B = \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| \leqq \mu(N)\}$$

and Δ is the usual Laplacian operator in $\mathbb{R}^3$. (Although for clarity we restrict our discussion to $\mathbb{R}^3$, the techniques described in this paper apply equally to $\mathbb{R}^n$ with $n > 2$.)

When p is less than the critical Sobolev exponent for $\mathbb{R}^3$, namely, $p_c \equiv 5$, the solutions of problem (1.2) may be studied by employing methods from the calculus of variations. Details of this study are given by Brezis and Nirenberg in [2]. When $p > 5$, however, much less is known about the solutions of problems (1.1), (1.2). The object of this paper is to show how ideas from the theory of dynamical systems may be used to obtain information about the structure of solutions of problem (1.1) for this range of p. We will establish the following theorem.

THEOREM 1.1. *Let $u(r)$ be a solution of problem* (1.1); *then there exists a value $\mu_\infty < \infty$ and a sequence N_n with $n \in \mathbb{N}$ such that $u(0) \equiv N_n \to \infty$ as $n \to \infty$ and $\mu(N_n) = \mu_\infty$.*

* Received by the editors August 24, 1987; accepted for publication (in revised form) November 18, 1988.
† Oxford University Computing Laboratory, Keble Road, Oxford OX1 3QD, United Kingdom.

This theorem implies, in particular, that problem (1.2) has an infinite number of positive solutions when $B \equiv B_\infty$ has radius μ_∞.

A proof of Theorem 1.1 was given originally by Budd and Norbury [3], but it is technical and requires careful use of methods from functional analysis. In this paper we use methods from the theory of dynamical systems to give an alternative proof of this result and to provide a deeper insight into the structure of the solution set. In particular, we use methods from Shilnikov's theory, the theory of normal forms, and a theorem due to Belitskii to obtain our result. Similar ideas have been used to study problems related to (1.1) with $p < p_c$ and some examples are given by Jones and Küpper [11] and by Jones, Küpper, and Plakties [12].

The function u^q in problem (1.1) may be replaced by a more general function $g(u)$ satisfying the three conditions (E1)-(E3) described in § 3. We may further extend the techniques described in this paper to study other "supercritical" systems also having nonlinearities that grow too rapidly for the usual techniques from the calculus of variations to be applicable. An example of such a system is the degenerate Laplacian equation

$$\nabla \cdot [|\nabla u|^{m-2}\nabla u] + u^p + u^q = 0, \tag{1.3}$$

with

$$p > [(m-1)n + m]/(n-m), \tag{1.4}$$

described in Ni and Serrin [16]. (Here n is the dimension of the space, and we note that problem (1.3) reduces to (1.1) if $m = 2$ and $n = 3$.)

2. Dynamical systems related to elliptic equations. The idea of using dynamical system methods to study problems similar to (1.1) was introduced by Fowler [7] with further examples given by Joseph and Lundgren in [13]. Recently, Jones and Küpper [11] have used such methods extensively to study the case $p < 5$; in [4] we examine the case $p \approx 5$. In the latter two cases, particular attention is given to special topological features of the phase plane associated with (1.1) that alter as p passes through the critical value $p_c = 5$.

We introduce the following variables (described originally by Fowler [7]).

Let $t = \operatorname{Ln} r$ and let

$$s(t) = r^\gamma, \quad a(t) = r^\alpha u(r), \quad b(t) = r^{1+\alpha} u_r(r), \tag{2.1}$$

where $\alpha = 2/(p-1)$ so that $\alpha < \frac{1}{2}$ when $p > 5$ and

$$\gamma = \frac{p-q}{p-1}. \tag{2.2}$$

(The introduction of the third variable $s(t)$ is crucial to the theory developed in this paper, as it allows us to transform the second-order ordinary differential equation (1.1) into a three-dimensional autonomous dynamical system. The original idea of using this extra variable is due to Jones and Küpper [11].)

Under this change of variables, problem (1.1) transforms into the following dynamical system:

$$\begin{aligned} \frac{da}{dt} &= \dot{a} = \alpha a + b, \\ \frac{db}{dt} &= \dot{b} = (\alpha - 1)b - a^p - s^2 h(a, s), \\ \frac{ds}{dt} &= \dot{s} = \gamma s, \end{aligned} \tag{2.3}$$

where $h(a, s) \equiv a^q$. Our choice of γ here ensures that all coefficients in (2.3) are C^∞ functions of a, b, and s if $a \neq 0$. We may, in fact, consider problems with a slightly more general nonlinearity than u^q, namely, $g(u)$. To retain the regularity of the coefficients of problem (2.3) we will insist that $g(u)$ satisfies the following condition:

$$s^{2p/(p-q)} g(as^{-2/(p-q)}) = s^2 h(a, s),$$

where the function $s^2 h(a, s)$ is $C^{1,1}$ in a and s if $a \neq 0$. It is this more general problem that we consider for the remainder of this paper. An example of such a function is $g(u) = ju^q + ku^{q'}$ with $q' < q$. The existence of a regular solution to problem (1.1) is guaranteed by the usual existence theory described, for example, by Smoller and Wasserman in [21]. Such a solution satisfies the conditions $u(r) \to N$, $u_r(r) \to 0$ as $r \to 0$ and $u(\mu) = 0$. From the form of the transformation (2.1) it is evident that the corresponding solution of problem (2.3) is a one-dimensional trajectory that leaves the origin $(a, b, s) = (0, 0, 0)$ when $t = -\infty$ and intersects the plane $a = 0$ when $t = \log \mu(N)$. Because the differential equation (1.1) is derived from an elliptic partial differential equation we may immediately deduce some properties of the corresponding solutions of problem (2.3).

LEMMA 2.1. (i) *If $N > 0$, then the solution trajectory of problem* (2.1) *corresponding to the solution of* (1.1) *initially lies in the quadrant $a > 0 > b$.*

(ii) *If $b = 0$, then $a\dot{b} < 0$.*

(iii) *If $ab > 0$, then $a\dot{a} \geqq 0$.*

(iv) *If $a = 0$, then $\dot{a}b > 0$.*

Proof. Observations (i) and (ii) are immediate deductions from the maximum principle (see Protter and Weinberger [19]), while (iii) and (iv) follow from the definitions of a and b. □

We now make a more detailed study of the system described by (2.3). If $p > 3$ the origin is an unstable saddle point with eigenvalues $\gamma, \alpha > 0$ and $(\alpha - 1) < 0$ with corresponding eigenvectors $(a, b, s) = (0, 0, 1)$; $(1, 0, 0)$ and $(1, -1, 0)$. A trajectory corresponding to a solution of (1.1) leaves the origin tangent to the eigenvector $(1, 0, 0)$. (It is especially interesting that the nature of the singularity at the origin changes precisely at the point $p = 3$, where the uniqueness proofs of Coffman [6] and McLeod and Serrin [15] do not apply to problems similar to (1.1). However, a recent result due to Kaper and Kwong [23] has extended the techniques above to the complete range $1 < p < (n+2)/(n-2)$ in $\mathbb{R}^n$.)

The remaining singular points of the system (2.3) are located at the two points

$$\mathbf{P}_\pm \equiv (a, b, s) = \pm(k, -\alpha k, 0), \qquad k^{p-1} = \alpha(1 - \alpha). \tag{2.4}$$

A local analysis of (2.1) at the points $\mathbf{P}_\pm$ shows that its linearisation about this point has eigenvalues

$$(\alpha - \tfrac{1}{2} \pm i\omega, \gamma) \quad \text{where } \omega^2 = 2p(p-3)(p-1)^{-2} - \tfrac{1}{4}. \tag{2.5}$$

When $3 < p < 5$ then $\alpha > \frac{1}{2}$ and the singular points are unstable spirals. However, they become stable spirals if we reverse the sign of t. This transformation in t allows us to use the methods described in this paper to study the solutions of problem (1.1) for $p < 5$ and $r \gg 1$. When $p = 5$ then $\alpha = \frac{1}{2}$, and the restriction of system (2.3) to the plane $s = 0$ leads to a Hamiltonian dynamical system. This structure is exploited by Budd in [4], where system (2.3) for the particular exponent range $(p - 5) \ll 1$ is studied as an example of a perturbed Hamiltonian system.

When $p>5$, α satisfies the inequality $0<\alpha<\frac{1}{2}$ and the singular point has two complex conjugate eigenvalues with real part $\alpha-\frac{1}{2}<0$ and an eigenvalue $\gamma>0$. This point is therefore of the saddle focus type. (For the more general degenerate Laplacian equation (1.3), if we take $m=2$ then the points $\mathbf{P}_{\pm}$ are spiral attractors if $2<n<10$, with a similar restriction on the value of n for other values of m.) We may now deduce the following results from the Stable Manifold Theorem (Guckenheimer and Holmes [9]).

LEMMA 2.2. *The point $\mathbf{P}_+$ has a two-dimensional stable manifold W_s and a one-dimensional unstable manifold $M(t)$. The function $M(t)$ is a C^∞ function of t and is tangent to the line $(k, -\alpha k, e^{\gamma t})$ as $t\to-\infty$. The set W_s is a subset of the plane $s=0$.*

Proof. This result follows immediately from the regularity of the coefficients in (2.3) and their behaviour as the term $(a-k, b+\alpha k, s)\to 0$. □

This special structure of the points $\mathbf{P}_{\pm}$ allows us to consider the use of solution methods similar to those developed by Shilnikov [20] when studying the solution structure of problem (1.1). We observe, however, that as $\dot{s}\geqq 0$ for all $t>0$, the trajectory $M(t)$ does not return to the point $\mathbf{P}_{\pm}$ as $t\to\infty$, and therefore we may not apply Shilnikov methods directly to problem (2.3). We will show later, however, that the topology of $M(t)$ plays an important role in the structure of the solutions of problem (2.3).

To proceed further, we now study the attracting set at the point $\mathbf{P}_+$, which lies in the plane $s=0$. A straightforward calculation shows that this plane in an invariant of the system (2.3) and that the restriction of this system to the plane gives the following autonomous dynamical system:

$$\dot{a}=\alpha a+b, \qquad \dot{b}=(\alpha-1)b-a^p. \tag{2.6}$$

To study this system we make the following calculation, described, for example, by Joseph and Lundgren [13]. Let the function $W(N, r)$ be the solution of the following ordinary differential equation problem:

$$\begin{aligned} &W_{rr}+\frac{2}{r}W_r+W^P=0,\\ &W(N,0)=N, \qquad W_r(0)=0. \end{aligned} \tag{2.7}$$

An application of the transformation (2.1) to the function $W(N, r)$ transforms (2.7) into the system (2.6). Let $(A_N(t), B_N(t))$ be the solution trajectory of problem (2.6) that corresponds to $W(N, r)$. It is significant that the effect of N in the transformation above does not alter the locus described by $(A_N(t), B_N(t))$ but merely alters its parameterisation with t. We show this by establishing the following lemma.

LEMMA 2.3. *Let $(A_N(t), B_N(t))$ be defined as above; then*

$$(A_N(t), B_N(t))=(A_1(t+\log N/\alpha), B_1(t+\log N/\alpha)).$$

Proof. It is well known that if $W(N, r)$ is the solution of problem (2.7), then

$$W(N, r)=NW(1, rN^{1/\alpha}). \tag{2.8}$$

Hence,

$$\begin{aligned} A_N(t)&=Ne^{\alpha t}W(1, e^{t+\log N/\alpha})\\ &=e^{\alpha(t+\log N/\alpha)}W(1, e^{t+\log N/\alpha})=A_1(t+\log N/\alpha). \end{aligned}$$

A similar calculation follows for $B_N(t)$. □

It is further shown by Chandrasekhar [5] that a solution trajectory of the system (2.6) corresponding to $W(N, r)$ leaves the origin when $t=-\infty$ and tends toward $\mathbf{P}_+$ as $t\to\infty$. Thus, if we define a curve $\mathbf{O}_+$ to be the locus described by $(A_N(t), B_N(t), 0)$,

then $\mathbf{O}_+$ lies in the stable manifold for the point $\mathbf{P}_+$ defined by system (2.3) while also being an unstable manifold for the origin.

Figure 1 shows the point $\mathbf{P}_+$ and gives a picture of the locus of $M(t)$ and of $\mathbf{O}_+$.

3. Applications of Shilnikov's theory. The original applications of the ideas of Shilnikov [20] lie in the examination of periodic orbits associated with a homoclinic orbit that leaves a stationary point $\mathbf{P}$ along its unstable manifold and that returns along a spiral trajectory in the attracting set of $\mathbf{P}$. Examples of these applications are given by Glendinning and Sparrow [8] and by Tresser [22]. Although this structure does not occur for the system (2.3) we may exploit specific features of the original elliptic system (1.1) to allow us to employ similar techniques to study problem (2.3). We now state the main result of this section.

THEOREM 3.1. *Let $u(r)$ be the solution of the ordinary differential equation problem*

$$u_{rr}+\frac{2}{r}u_r+u^p+g(u)=0, \tag{3.1}$$
$$u(0)=N, \qquad u_r(0)=0.$$

Furthermore, we suppose that the following conditions hold.

(E1) *The function $g(u)$ satisfies the following regularity condition: There is a positive value $q<p$ such that*

$$s^{2p/(p-q)}g(as^{-2/(p-q)})=s^2h(a,s),$$

where the function $s^2h(a,s)$ is $C^{1,1}$ as $s\to 0$ $(a\neq 0)$ and is C^2 if $s\neq 0$.

(E2) *Let $V(N,r)$ be a solution of the scaled equation*

$$V_{rr}+\frac{2}{r}V_r+V^p+N^{-p}g(NV)=0, \tag{3.2}$$
$$V(0)=1, \qquad V_r(0)=0$$

and let $W(1,r)$ be a solution of problem (2.6). Then for each fixed $r>0$

$$|(V(N,r),V_r(N,r))-(W(1,r),W_r(1,r))|<C(r)N^{q-p}$$

as $N\to\infty$, where $C(r)$ does not depend on N.

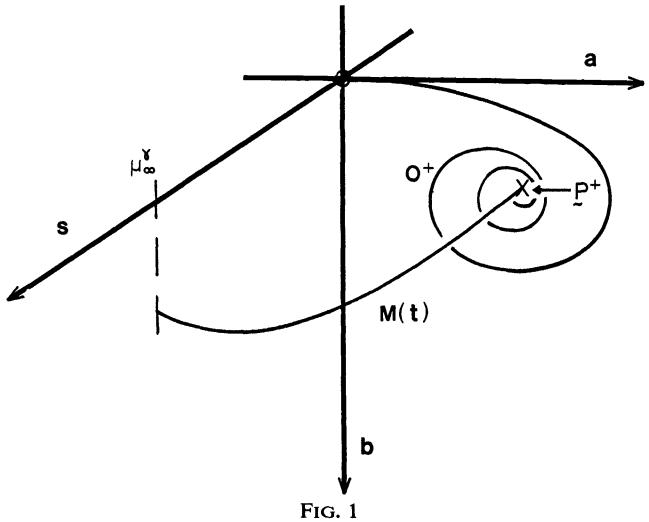


FIG. 1

(E3) *Let $M(t)$ be the unstable manifold of $\mathbf{P}_+$ constructed in § 2; then $M(t)$ intersects the plane $a=0$ when $s=(\mu_\infty)^\gamma<\infty$.*

Then, if $\mu(N)$ is defined to be the first positive zero of $u(r)$, as $N\to\infty$,

$$\mu(N)=\mu_\infty+EK(N)\cos(\omega \operatorname{Ln} N/\alpha+F)(1+o(1))+O(K(N)^2), \tag{3.3}$$

where $K(N)=N^{-(p-5)/4}$ and E, F are constants.

In § 4 we show that each of (E1)-(E3) is satisfied if $g(u)=u^q$ and $q<3$.

Proof. To prove this result we will define a map from N to a set Σ_0 in the phase space (a, b, s) defined by the transformation (2.1) and we will then employ Shilnikov's theory to map Σ_0 to a neighbourhood of the intersection of $M(t)$ with the plane $a=0$. We define Σ_0 as follows.

Let $\mathbf{O}_+$ be the curve constructed in § 2 and let (e,f) be a point lying on $\mathbf{O}_+$ in a neighbourhood of the point $\mathbf{P}_+$. We suppose that $(e,f)=(A_1(t_1), B_1(t_1))$ and we set $t_N=t_1-\log N/\alpha$. Let the curve $(a_N(t), b_N(t), s_N(t))$ be the solution trajectory of system (2.2) corresponding to the solution of problem (3.1). We define the set Σ_0 by

$$\Sigma_0=\{(a_N(t_N)b_N(t_N), s_N(t_N)): N>0\}.$$

LEMMA 3.2. *The set Σ_0 is a one-dimensional submanifold of $\mathbb{R}^3$. As $N\to\infty$, then*

$$\begin{aligned}&|(a_N(t_N)-e, b_N(t_N)-f)|<2C(\exp(t_1))N^{q-p},\\ &s_N(t_N)=\exp(\gamma t_1)N^{(q-p)/2}.\end{aligned} \tag{3.4}$$

Thus Σ_0 approaches the point $(e, f, 0)$ and is ultimately tangent to the line $(a, b)=(e, f)$.

Proof. The topological structure of Σ_0 as a subset of $\mathbb{R}^3$ follows from the analytic dependence of $u(r)$ on N. To establish (3.4) we make use of the condition (E2). A simple calculation shows that $V(N, r)=N^{-1}u(rN^{-1/\alpha})$, and hence we may deduce from (E2) that if r is fixed, then

$$r^\alpha|N^{-1}u(rN^{-1/\alpha})-W(1, r)|<r^\alpha C(r)N^{q-p} \quad \text{as } N\to\infty.$$

We now put $r=\exp(t_1)$, which implies that $r^\alpha N^{-1}=\exp(t_N)$. Thus $r^\alpha N^{-1}u(rN^{-1/\alpha})=e^{\alpha t_N}u(e^{t_N})=a_N(t_N)$ and $r^\alpha w(1, r)=e$. This gives the bound for $a_N(t_N)$ and a similar calculation describes the behaviour of $b_N(t_N)$.

Finally, $s_N(t_N)=\exp(\gamma t_N)=\exp(\gamma t_1)N^{-\gamma/\alpha}$. □

We will now construct a map Φ from the set Σ_0 to a set Σ_1 lying in the plane $a=0$ and including a neighbourhood of the point Q, where

$$Q=M(t)\cap\{a=0\},$$

and where the existence of the point Q is guaranteed by our assumption (E3).

Let $\mathbf{Z}$ be a point lying in Σ_0 and suppose that $\mathbf{T}(t)$ is the solution trajectory of problem (2.3) that intersects $\mathbf{Z}$. We define $\Phi(\mathbf{Z})$ such that $\Phi\colon \Sigma_0\to\Sigma_1$ by

$$\Phi(\mathbf{Z})=\mathbf{T}(t)\cap\Sigma_1.$$

Our main geometrical result describing the map Φ is given as follows.

LEMMA 3.3. *The image of Σ_0 under the map Σ_1 is a logarithmic spiral lying in Σ_1 centered on Q.*

Proof. To prove this result we proceed as in the proof of Shilnikov's theorem [20] using a presentation based on that given by Guckenheimer and Holmes [9]. We decompose Φ into two operators Φ_I and Φ_0 such that Φ_I maps Σ_0 to a set Σ_ε lying in a neighbourhood of $\mathbf{P}_+$ and Φ_0 maps Σ_ε to Σ_1.

If we set $c(t) = a(t) - k$ and $d(t) = b(t) + \alpha k$ then close to $\mathbf{P}_+$, system (2.3) has the following form:

$$\begin{aligned} \dot{c} &= \alpha c + d, \\ \dot{d} &= \alpha(1-\alpha)c + (\alpha - 1)d + R(s, c, d), \\ \dot{s} &= \gamma s. \end{aligned} \tag{3.5}$$

A simple calculation using the smoothness of the function $h(a, s)$ defined by the assumption (E1) shows that the function $R(s, c, d)$ is $C^{1,1}$ in a neighbourhood of the point $(c, d, s) = (0, 0, 0)$. Furthermore $R(\mathbf{0}) = DR(\mathbf{0}) = 0$ (although R has nonvanishing terms of second order in c, d, and s).

To simplify our calculations on system (3.5) we introduce a local change of coordinates to eliminate the nonlinear term $R(s, c, d)$. The existence of a homeomorphism with this property is guaranteed by the Hartman-Grobman theorem [9]. However, to obtain the results needed in this paper, we require our map to be a C^1 diffeomorphism whose existence is guaranteed by the following result due to Belitskii [1].

THEOREM 3.4 (Belitskii). *Let* $\mathbf{x}$ *satisfy the differential equation*

$$\dot{\mathbf{x}} = L\mathbf{x} + R(\mathbf{x}),$$

where L *is a linear operator,* $R(\mathbf{x}) \in C^{1,1}$, *and* $R(0) = DR(0) = 0$. *Furthermore, suppose that the eigenvalues* λ_i *of* L *satisfy the nonresonance condition*

$$\lambda_i \neq \lambda_j + \lambda_k \tag{3.6}$$

if $\operatorname{Re}(\lambda_j) < 0 < \operatorname{Re}(\lambda_k)$. *Then there exists a* C^1 *diffeomorphism* $T: x \to y$ *such that*

$$T(0) = 0, \quad DT(0) = I, \quad \dot{y} = L\mathbf{y}.$$

Proof. See [1] and the discussion of normal forms in Chapter 3 of [9]. □

LEMMA 3.5. *Let* L *be the linear operator in system* (3.5); *then condition* (3.6) *is satisfied.*

Proof. The eigenvalues of L are given by (2.5) and have the values $(5-p)/2(p-1) \pm i\omega$ and $(p-q)/(p-1)$. Condition (3.6) follows from the observation that $\omega \neq 0$. The regularity of the coefficients required for an application of the theorem is immediate from condition (E1). □

LEMMA 3.6. *Let* T *be the map constructed in Theorem* 3.4 *such that* $T: (c, d, s) \to (C, D, S)$. *In a neighbourhood* $\mathfrak{M}$ *of* $\mathbf{P}_+$ *the system* (3.5) *takes the following form*:

$$\begin{aligned} \dot{C} &= \alpha C + D, \\ \dot{D} &= \alpha(1-\alpha)C + (\alpha - 1)D, \\ \dot{S} &= \gamma S. \end{aligned} \tag{3.7}$$

Furthermore, T *maps the plane* $\mathfrak{M} \cap \{s = 0\}$ *into the plane* $\{S = 0\}$.

Proof. Result (3.7) follows immediately from the definition of T. Now if $\mathfrak{M}$ is sufficiently small, any point in $\mathfrak{M} \cap \{s = 0\}$ lies on a stable manifold of the origin. Consequently its image under T must also lie on a stable manifold of the origin in the transformed coordinates and hence lies in $\{S = 0\}$. □

It is evident from inspection of system (3.7) that the image $\hat{M}(t)$ of the unstable manifold $M(t)$ of the point $\mathbf{P}_+$ is locally the curve $(C, D, S) = (0, 0, e^{\gamma t})$. Now let $\hat{\Sigma}_\varepsilon$ be a set lying in the image of $\mathfrak{M}$, centred on the point $(0, 0, \varepsilon)$ and lying in the plane $S = \varepsilon$. Furthermore, let us take the point $(e, f, 0)$ sufficiently close to $\mathbf{P}_+$ so that if N_0 is sufficiently large, a subset of Σ_0 defined by taking $N > N_0$ lies in $\mathfrak{M}$. We define $\hat{\Sigma}_0$

to be the image of Σ_0 under T and construct a map Φ_I from $\hat{\Sigma}_0$ to $\hat{\Sigma}_\varepsilon$ as follows. Let $\hat{\mathbf{Z}} \in \hat{\Sigma}_0$ and let $\hat{u}$ be the solution trajectory of (3.7) that contains $\hat{\mathbf{Z}}$. We define Φ_I by

$$\Phi_I(Z) \equiv (C_\varepsilon, D_\varepsilon) = \hat{u} \cap \hat{\Sigma}_\varepsilon.$$

LEMMA 3.7. *Let $\hat{Z} = (x, y, S)$ with $S < \varepsilon$; then*

$$\Phi_I(\hat{Z}) = e^{(\alpha - 1/2)\delta}(x \cos \omega\delta - y \sin \omega\delta, x \sin \omega\delta + y \cos \omega\delta), \tag{3.8}$$

where $\delta = \log((\varepsilon s)^{-1/\gamma})$ and ω is defined by (2.5).

Proof. This result follows immediately on explicit solution of the linear system (3.7). □

We now examine the nature of the set $\hat{\Sigma}_0$.

LEMMA 3.8. *Let the point $(E, F, 0)$ be the image of the point $(e - k, f + \alpha k, 0)$ under the map T. Furthermore, let $u_N \equiv (a_N(t), b_N(t), s_N(t))$ be the solution trajectory of* (2.3) *corresponding to the solution of problem* (3.1). *Then the point $\hat{\mathbf{Z}}_N$ where the image of this trajectory under T intersects $\hat{\Sigma}_0$ is given by*

$$\hat{\mathbf{Z}} = (E, F, \exp(\gamma t_1) N^{(q-p)/2}) + O(mN^{(q-p)/2}) + O(N^{q-p}), \tag{3.9}$$

where $\|m\|$ may be taken arbitrarily small by taking (E, F) sufficiently close to $(0, 0)$.

Proof. To prove this result we use the fact that the map T is C^1 near the origin and $DT(0) = I$. Then

$$u_N \cap \Sigma_0 \equiv \mathbf{Z}_N = (e - k, f + \alpha k, 0) + \ell_N,$$

where

$$\ell_N = (O(N^{q-p}), O(N^{q-p}), \exp(\gamma t_1) N^{(q-p)/2}).$$

(Here the estimate for ℓ_N follows from (3.4).)

From the continuity of DT we may deduce that $DT(e - k, f + \alpha k, 0) = I + m$, where $\|m\|$ may be taken arbitrarily small by taking $(e - k, f + \alpha k)$ sufficiently close to $(0, 0)$. It thus follows from the mean value theorem that $\hat{\mathbf{Z}}_N = TZ_N$ has the form (3.9). □

We may now explicitly calculate the image of $\hat{Z}_N$ under the action of Φ_I in terms of the parameter N.

LEMMA 3.9. *Let u_N be constructed as in Lemma* 3.8; *then the point $\hat{\mathbf{y}}_N$, where the image of u_N under T intersects $\hat{\Sigma}_\varepsilon$, is given by*

$$\begin{aligned} &\hat{\mathbf{y}}_N \equiv \hat{\mathbf{x}}_N + (0, 0, \varepsilon), \text{ where } \hat{\mathbf{x}}_N = (A_1 K(N)[\cos(A_2 + \omega \log(DN^{1/\alpha})), \\ &\sin(A_2 + \omega \log(DN^{1/\alpha})] * ((1 + O(m)), 0) \text{ as } N \to \infty, \end{aligned} \tag{3.10}$$

where $K(N) = N^{-(p-5)/4}$ with A_1, A_2, and D constant.

Proof. We combine the results (3.8) and (3.9). From (3.9) we see that the value of δ in (3.8) is given by

$$\delta = \log(DN^{1/\alpha}(1 + O(m)), \qquad N \to \infty,$$

where D is a suitable constant. Thus

$$e^{(\alpha - 1/2)\delta} = A_1 N^{-(p-5)/4}(1 + O(m)). \tag{3.11}$$

Similarly, $x \cos \omega\delta = [E + O(N^{(q-p)/2})] \cos \omega\delta$, with a similar result for y. Putting these values together and collecting the error terms of order m, we deduce (3.10). □

Thus as N increases, the point $\hat{\mathbf{y}}_N$ traces out a logarithmic spiral ζ_ε lying in $\hat{\Sigma}_\varepsilon$ and centred on $(0, 0, \varepsilon)$.

LEMMA 3.10. *Let $\mathbf{y}_N$ be the pre-image of the point $\hat{\mathbf{y}}_N$ under the map T so that $T(\mathbf{y}_N) = \hat{\mathbf{y}}_N$. Then*

$$\mathbf{y}_N = T^{-1}(0, 0, \varepsilon) + \mathbf{x}_N, \qquad \text{where } \mathbf{x}_N = \hat{\mathbf{x}}_N(1 + O(m)). \tag{3.12}$$

Proof. We may deduce from Theorem 3.4 that the map T is a diffeomorphism. The result (3.12) then follows by linearising the inverse of T about the point $(0, 0, \varepsilon)$. We note that if ε is small the linearisation of T^{-1} is close to the identity so that we may take $\|T^{-1} - I\| = O(m)$. □

We will now construct a map Φ_0 from the pre-image Σ_ε of the set $\hat{\Sigma}_\varepsilon$ to the set Σ_1 by computing the flow of solutions of problem (2.3) that have initial data lying in Σ_ε. Let $\mathbf{Q}$ be the intersection of $M(t)$ with the plane $a = 0$ (the existence of which is guaranteed by assumption (E3)) and let $\mathbf{Q}_\varepsilon$ be the intersection of $M(t)$ with Σ_ε. Thus $\mathbf{Q}_\varepsilon = T^{-1}(0, 0, \varepsilon)$. The flow from $\mathbf{Q}_\varepsilon$ along $M(t)$ to $\mathbf{Q}$ is nonsingular and thus from the regularity condition (E1) the flow map Φ_0 is a diffeomorphism from the neighbourhood of $\mathbf{Q}_\varepsilon$ in $\hat{\Sigma}_\varepsilon$ to a neighbourhood of $\mathbf{Q}$ in the plane $a = 0$. Hence the image of the logarithmic spiral ζ_ε under the action of Φ_0 is a spiral lying in the plane $a = 0$. This proves Lemma 3.3. □

To complete the proof of Theorem 3.1 we consider the form of the spiral ζ. As the map Φ_0 is a C^2 diffeomorphism we may locally linearise it about the point $\mathbf{Q}_\varepsilon$. Thus the image of the point $\mathbf{Q}_\varepsilon + \mathbf{x}_N$, when $\|x_N\|$ is sufficiently small, is given by

$$\Phi_0(\mathbf{Q}_\varepsilon + \mathbf{x}_N) = \mathbf{Q} + D\Phi_0 \mathbf{x}_N + O(\|x_N\|^2) \tag{3.13}$$

$$= \mathbf{Q} + D\Phi_0 \hat{\mathbf{x}}_N (1 + O(m) + O(K(N))), \tag{3.14}$$

where the point $\mathbf{Q} + D\Phi_0 \mathbf{x}_N + O(\|x_N\|^2)$ lies in the plane $a = 0$. The s coordinate of this point thus gives us the value of $\mu(N)$ defined in Theorem 3.1. Furthermore, the coordinates of the point $\mathbf{Q}$ are given by

$$\mathbf{Q} = (0, \mu_\infty^{\alpha+1} M'(\mu_\infty), \mu_\infty^{\gamma}).$$

To obtain (3.3) we combine the two expressions (3.10) and (3.14). The values of the constants E and F in (3.3) follow from the coefficients of the linear operator $D\Phi_0$ described by (3.14).

This concludes the proof of Theorem 3.1. □

We indicate in Fig. 2 the main steps in the proof of this theorem, and in Fig. 3 we further show a graph of $\mu(N)$ obtained from the expression (3.3). Figure 3 shows a very strong similarity to Fig. 3.4 in [9]. This observation reinforces the link indicated

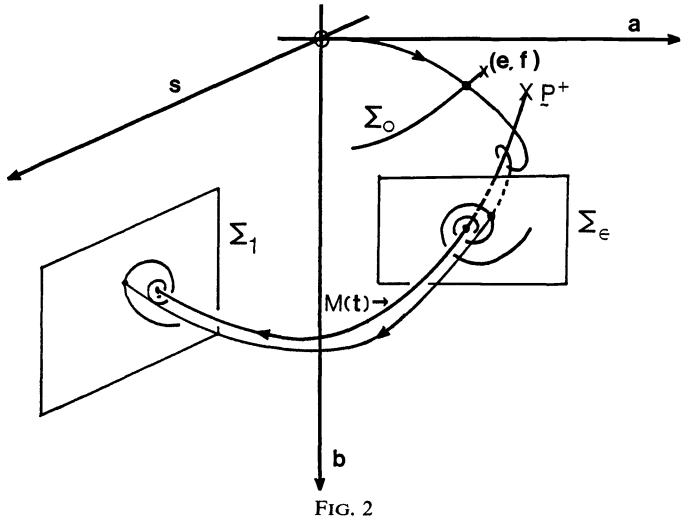


FIG. 2

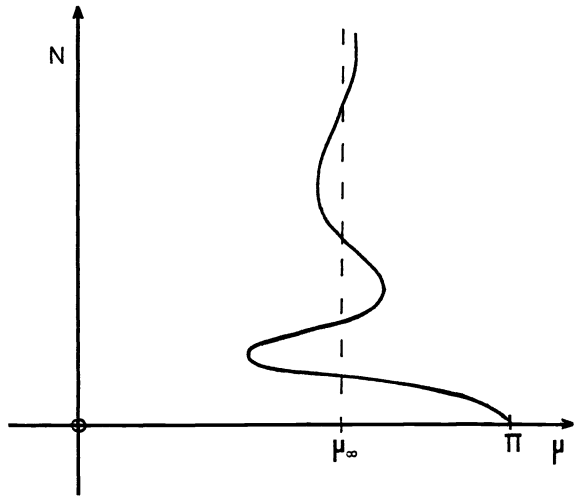


FIG. 3

in this paper between the structure of the homoclinic orbits studied by Shilnikov and the elliptic equations considered here.

4. Applications to elliptic equations. To apply Theorem 3.1 to elliptic equations of the form (1.1) we must establish that conditions (E1)-(E3) hold for system (3.1) if $g(u) \equiv u^q$. Initially we will establish condition (E2), as (E1) is automatically satisfied for the nonlinearity u^q.

LEMMA 4.1. *Let $V(N, r)$ be a solution of problem* (3.2) *and let $g(u)$ satisfy the condition that $g(u)/u^q$ is bounded above as $u \to \infty$. Then there is a function $K(r)$ bounded independently of N such that*

$$|(V(N, r) - W(1, r), V_r(N, r) - W_r(N, r))| < N^{q-p}K(r).$$

Proof. We may observe from the maximum principle that the function $V(N, r)$ is bounded above by 1 on the interval on which it is positive. We may thus deduce from our assumption on the growth of the function $g(u)$ that the function $N^{-p}g(NV)$ is bounded by $N^{q-p}k(V)$, where $k(V)$ is bounded independently of N as $N \to \infty$.

We define L to be the following Volterra integral operator:

$$Lf(s) = \int_0^s t(1-(t/s))f(t)\,dt \equiv \int_0^s m(t, s)f(t)\,dt.$$

Problem (3.1) may then be put into the following form:

$$V(s) = 1 + L[V^p + N^{q-p}k(V)].$$

Now let $W(r)$ be the solution of problem (2.7) with $N = 1$. It follows that

$$W(r) = 1 + L[W^p].$$

If we now set $x(r) = W(r) - V(r)$, then $x(r)$ satisfies the following equation:

$$x(r) = L[(V+x)^p - V^p - N^{q-p}k(V)]. \tag{4.1}$$

By the Mean Value Theorem we may deduce that

$$(V+x)^p - V^p = p[V + \theta x]^{p-1}x,$$

where $0<\theta<1$. Thus from the maximum principle we can see that $|(V+\theta x)|<1$ and hence

$$|x(s)| \leqq L[p|x|+N^{5-p}|k(V)|]. \tag{4.2}$$

As the kernel $m(r, s)$ is bounded we may deduce from Gronwall's Lemma [10] that there is a function $K(r)$ bounded independently of N such that

$$|x(r)| \leqq N^{(5-p)}K(r). \tag{4.3}$$

A similar bound for $|x_r(r)|$ then follows on differentiating the identity (4.2) and substituting the bound given in (4.3). This establishes condition (E2). (A very similar calculation can also be made to study the degenerate Laplacian problem (1.3).) □

To complete the proof of Theorem 1.1 we must establish condition (E3), which entails a global characterisation of the manifold $\mathbf{M}(t)$. A simple calculation shows that such a solution of the problem (2.3) corresponds to a singular solution $M(r)$ of the differential equation problem (1.1). This solution satisfies the conditions

(4.4) $r^{\alpha}M(r)\to k$ and $r^{\alpha+1}M_r(r)\to -\alpha k$ as $r\to 0$, where k is defined by the identity (2.4).

We may now appeal to the following theorem, established by Ni and Serrin [17].

THEOREM 4.2. *Let $M(r)$ be the solution of problem* (1.1) *that satisfies the singular initial conditions described in* (4.4). *Furthermore, suppose that for some $q<3$ the function $g(u)$ has the form*

$$u^p+g(u)=\lambda u^q+h(u) \quad \lambda>0,$$

where

$$h(u)>0 \quad \text{if } u>0.$$

If these conditions hold, then there is a value $\mu_\infty<\infty$ such that $M(\mu_\infty)=0$ and thus (E3) is satisfied. □

The condition $q<3$ stated above is important to the proof of Theorem 4.2 as it rules out the possibility of a "ground-state" solution of problem (1.1), namely, a solution that tends to zero as $r\to\infty$ but is everywhere positive. It is not clear, however, what happens in the case $3\leqq q<5$. A recent result due to Lin and Ni [14] shows that there does exist a ground-state solution of problem (1.1) if $p>5$ and $q=(p+1)/2$, although this does not necessarily imply that condition (E3) is not true in this case. Numerical calculations seem to imply that (E3) remains true for all q such that $3<q<5$. For $q\geqq 5$ an application of Pohozaev's identity [18] implies that condition (E3) cannot be satisfied.

Again a similar calculation can be made for the singular solutions of the degenerate Laplacian equation.

Having established identity (3.3) for the solutions of problem (1.1), we may now establish Theorem 1.1 by choosing values of N such that the function $G(N)$ vanishes where

$$G(N)\equiv \cos(\omega \operatorname{Ln} N/\alpha+F)(1+o(1))+O(K(N)).$$

Because $K(N)$ tends to zero as N tends to infinity, the existence of such a sequence is guaranteed. This proves Theorem 1.1. □

Acknowledgments. I am grateful to Professor C. Jones, Dr. J. Carr, Dr. P. Glendinning, and the referees for some remarks that helped to clarify an earlier version of this paper.

REFERENCES

[1] G. BELITSKII, *Functional equations and conjugacy of local diffeomorphisms of a finite smoothness class*, Functional Anal. Appl., 7 (1973), pp. 268-277.

[2] H. BREZIS AND L. NIRENBERG, *Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents*, Comm. Pure Appl. Math., 36 (1983), pp. 437-478.

[3] C. J. BUDD AND J. NORBURY, *Semilinear elliptic equations with supercritical growth rates*, J. Differential Equations, 68 (1987), pp. 169-197.

[4] C. J. BUDD, *Semilinear elliptic equations with near critical growth rates*, Proc. Roy. Soc. Edinburgh Sect. A, 107 (1987), pp. 249-270.

[5] S. CHANDRASEKHAR, *An Introduction to the Study of Stellar Structure*, Dover, New York, 1939.

[6] C. COFFMAN, *On the positive solution of boundary value problems for a class of nonlinear differential equations*, J. Differential Equations, 3 (1967), pp. 92-111.

[7] R. H. FOWLER, *Further studies of Emden's and similar differential equations*, Quart. J. Math., 2 (1931), pp. 259-288.

[8] P. GLENDINNING AND C. SPARROW, *Local and global behaviour near homoclinic orbits*, J. Statist. Phys., 35 (1984), pp. 645-695.

[9] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, Berlin, New York, 1983.

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

[11] C. JONES AND T. KÜPPER, *On the infinitely many solutions of a semilinear elliptic equation*, SIAM J. Math. Anal., 17 (1986), pp. 803-835.

[12] C. JONES, T. KÜPPER, AND H. PLAKTIES, *A shooting argument with oscillation for semilinear elliptic radially symmetric equations*, Proc. Roy. Soc. Edinburgh Sect. A, 108 (1988), pp. 165-180.

[13] D. JOSEPH AND T. LUNDGREN, Arch. Rational Mech. Anal., 49 (1973), pp. 241-269.

[14] C.-S. LIN AND W.-M. NI, *A counterexample to the nodal domain conjecture and a related semilinear elliptic equation*, Proc. Amer. Math. Soc., 102 (1988), pp. 271-277.

[15] K. MCLEOD AND J. SERRIN, *Uniqueness of solutions of semilinear Poisson equations*, Proc. Nat. Acad. Sci. U.S.A., 78 (1981), pp. 5692-6595.

[16] W.-M. NI AND J. SERRIN, *Existence and non-existence theorems for ground states of quasilinear* PDEs, University of Minnesota Mathematics Report, University of Minnesota, Minneapolis, MN, 1984, pp. 84-150.

[17] ———, *Non-existence theorems for singular solutions of quasilinear* PDEs, Comm. Pure Appl. Math., 39 (1986), pp. 379-399.

[18] S. POHOZAEV, *Eigenfunctions of the equation* $\Delta u + \lambda f(u) = 0$, Soviet Math. Dokl., 5 (1965), pp. 1408-1411.

[19] M. PROTTER AND H. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[20] L. SHILNIKOV, *A case of the existence of a denumerable set of periodic motions*, Soviet Math. Dokl., 6 (1965), pp. 163-166.

[21] J. SMOLLER AND A. WASSERMAN, *Symmetry breaking from positive solutions of semilinear elliptic equations*, in Proc. Dundee Conference on Differential Equations, Springer-Verlag, Berlin, New York, 1984.

[22] C. TRESSER, *About some theorems by L. P. Shilnikov*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 40 (1984), pp. 441-461.

[23] H. G. KAPER AND M. K. KWONG, *Uniqueness of non-negative solutions of a class of semi-linear elliptic equations*, in Proc. Berkeley Symposium on Nonlinear Diffusion Equations and their Equilibrium States, W.-M. Ni, L. A. Peletier, and J. Serrin, eds., Springer-Verlag, Berlin, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1081-1094, September 1989

004

ON A HYPERBOLIC QUENCHING PROBLEM IN SEVERAL DIMENSIONS*

RICHARD A. SMITH†

Abstract. Let $\phi \in C^1(-\infty, M)$ be nonnegative and increasing and satisfy $\lim_{u \to M^-} \phi(u) = \infty$. The problem $u_{tt} = \Delta_n u + \varepsilon\phi(u)$ in $D \times (0, T)$, $u = 0$ on $\partial D \times (0, T)$, $u(x, 0) = u_0(x)$, and $u_t(x, 0) = v_0(x)$ in D, is shown to have a unique local continuous solution for $\varepsilon > 0$ sufficiently small in dimensions $n = 1, 2, 3$ under appropriate assumptions on ϕ, u_0, v_0, and ∂D. The solution u can be continued as long as $u < M$. A potential well theory is shown to be unobtainable for this problem in the Sobolev space $H_0^1(D)$ for $n \geqq 2$; however, an a priori inequality for solutions guarantees global existence via energy considerations. Numerical evidence is given indicating that such an a priori inequality is sometimes satisfied by solutions when $n \geqq 2$.

Key words. hyperbolic partial differential equations, quenching

AMS(MOS) subject classification. 35

1. Introduction.

Let D be an open, bounded subset of $\mathbb{R}^n$ with boundary ∂D. Let ϕ: $(-\infty, M) \to (0, \infty)$, $M > 0$, be continuously differentiable, monotone increasing, convex, and satisfy $\lim_{u \to M^-} \phi(u) = \infty$; and let $\varepsilon > 0$. In this paper the following initial-boundary value problem is considered:

$$\begin{aligned} &u_{tt} = \Delta_n u + \varepsilon\phi(u) \quad \text{in } D \times (0, T), \\ &u = 0 \quad \text{on } \partial D \times (0, T), \\ &u(x, 0) = u_0(x), \qquad u_t(x, 0) = v_0(x) \quad \text{in } D, \end{aligned} \tag{An}$$

where Δ_n denotes the n-dimensional Laplacian.

Let (A1) denote problem (An) with $D = (0, 1)$. When $\phi(u) = 1/(M - u)$, a solution of (A1) has a physical interpretation as describing the motion of a wire composed of a magnetic material and carrying an electric current in the presence of another current-carrying wire [8]. Chang and Levine [2] have shown that for suitably regular initial data, problem (A1) has a unique local piecewise C^2 solution u that can be continued as long as $u < M$. They have also established the existence of numbers $\varepsilon_1 > \varepsilon_0 > 0$ such that

(a) If $\varepsilon \geqq \varepsilon_1$, then for some finite $T > 0$,

$$\lim_{t \to T^-} \left(\sup_{0 < x < 1} u(x, t) \right) = M;$$

hence one of u_{tt}, u_{xx} becomes infinite on $[0, 1] \times [0, T)$.

(b) If $0 < \varepsilon \leqq \varepsilon_0$, and the initial data u_0, v_0 are appropriately restricted, there is a $\delta = \delta(\varepsilon) > 0$ such that

$$|u(x, t)| \leqq M - \delta \quad \text{on } [0, 1] \times [0, \infty).$$

Note that by applying to (A1) the change of scale $x' = Lx$, $t' = Lt$, $\varepsilon = L^2$, we obtain (A1) with ε replaced by 1 and x' varying between 0 and L. Results (a) and (b) assert, therefore, that global solutions do not exist for long wires, but do exist for short wires.

* Received by the editors May 8, 1986; accepted for publication (in revised form) October 31, 1988.

† Exxon Production Research Company, P.O. Box 2189, N-225, Houston, Texas 77252-2189. This research was supported by Air Force Office of Scientific Research grant AFOSR 84-0252.

If u behaves as in (a), it is said to *quench* in finite time. Speaking loosely, we say that a solution of some evolutionary problem quenches in finite (or infinite) time T if some norm of the solution remains bounded, while some norm of one of its derivatives becomes unbounded, on the interval $[0, T)$ [7].

For space dimensions $n \geqq 2$ and ε sufficiently large, solutions of problem (An) also quench in finite time [2]. However, the proof of (b) in [2] relies strongly on the inequality

$$4u^2(x, t) \leqq \int_0^1 |u_x(x, t)|^2 \, dx, \tag{1.1}$$

which guarantees the imbedding of $H_0^1(0, 1)$ into $C([0, 1])$. In general, no such imbedding $H_0^1(D) \to C(\bar{D})$, or even $H_0^1(D) \to L_\infty(D)$, is possible in higher dimensions, and the question of existence of global solutions of (An) for $n \geqq 2$ remains open.

If instead of (An) we consider the abstract problem

$$\text{(B)} \qquad \begin{aligned} &\frac{d^2u}{dt^2} + Au = \varepsilon\phi(u), \qquad 0 < t < \infty, \\ &u(0) = u_0 \in V, \qquad u'(0) = v_0 \in L_2(D), \end{aligned}$$

where $V \subseteq L_2(D)$ imbeds in $L_\infty(D)$ and A is an operator of elliptic type mapping V into its dual, then a global existence theorem for ε sufficiently small may be proved (see Levine and Smiley [9]). Their results apply, for example, to solutions of (B) when D is the interior of a rectangle in $\mathbb{R}^2$, $u(x, t) = \Delta_2 u(x, t) = 0$ on $\partial D \times [0, \infty)$, A is the biharmonic operator Δ_2^2, and $V = H^2(D) \cap H_0^1(D)$.

Acker and Walter [1] have proved a higher-dimensional global existence theorem for small ε for solutions of

$$\text{(C)} \qquad \begin{aligned} &u_t = \Delta_n u + \varepsilon\phi(u) && \text{in } D \times (0, T), \\ &u = 0 && \text{on } \partial D \times (0, T), \\ &u(x, 0) = u_0(x) && \text{in } D. \end{aligned}$$

Their proof relies on consequences of the maximum principle for parabolic problems, which are available only in much weaker forms for hyperbolic problems. Hyperbolic problems in which the driving term $\varepsilon\phi(u)$ appears in a boundary condition instead of in the differential equation have also been studied [8], but the question of global existence of solutions in space dimensions higher than 1 also remains unanswered. For a comprehensive survey of the literature on quenching, see Levine [7].

In this paper, problem (An) is shown to have a unique local continuous solution u in low dimensions for small ε under appropriate assumptions on ϕ, u_0, v_0, and ∂D that can be continued as long as $u < M$. It is also shown that for any $\varepsilon > 0$, there exists no potential well about any equilibrium solution of (An), so that a proof of global existence along the lines of Sattinger [11] is not possible. Nevertheless, an a priori inequality for solutions of (An) similar to (1.1) is shown, via energy considerations, to guarantee global existence. Numerical evidence is given suggesting that such an a priori inequality is sometimes satisfied by solutions of (An).

2. Theoretical considerations. Local continuous solutions of (An) for $n = 2, 3$ are obtained by applying the abstract theory of Reed [10] to an appropriately modified problem.

In this section ϕ' will be assumed to be bounded and uniformly Lipschitz continuous on intervals of the form $(-\infty, M-\delta]$, $\delta>0$. For $0<\delta<M$ define

$$\phi_\delta(u)=\begin{cases}\phi(u), & u\leqq M-\delta,\\ \phi(M-\delta/2), & u\geqq M-\delta/2.\end{cases}$$

Then by suitably defining ϕ_δ on the interval $(M-\delta, M-\delta/2)$, we may arrange that $\phi_\delta\in C^1(\mathbb{R})$ and ϕ' is bounded and uniformly Lipschitz continuous on $\mathbb{R}$. Let (An, δ) represent problem (An) with ϕ replaced by ϕ_δ.

It is assumed that problem (An) has a stationary solution $f\in C^2(\bar{D})$, which is analytic in D and satisfies

$$\Delta_n f+\varepsilon\phi(f)=0 \quad \text{in } D,$$
$$f=0 \quad \text{on } \partial D.$$

Such is indeed the case when, e.g., D is a ball in $\mathbb{R}^n$ and $\phi(u)=(M-\alpha u)^\beta$, α, $\beta<0$ (see Joseph and Lundgren [5]).

Applying the transformation $\bar{u}=u-f$ to problem (An, δ), we obtain the problem

$$\begin{aligned}&\bar{u}_{tt}=\Delta_n\bar{u}+\varepsilon\psi_\delta(x,\bar{u}) \quad \text{in } D\times(0,T),\\ &\bar{u}=0 \quad \text{on } \partial D\times(0,T),\\ &\bar{u}(x,0)=u_0(x)-f(x), \quad \bar{u}_t(x,0)=v_0(x) \quad \text{in } D,\end{aligned}\tag{2.1}$$

where $\psi_\delta(x,u)\equiv\phi_\delta(u+f(x))-\phi_\delta(f(x))$. By setting

$$\bar{v}=\bar{u}_t, \quad \eta=\begin{pmatrix}\bar{u}\\ \bar{v}\end{pmatrix}, \quad \eta_0=\begin{pmatrix}u_0-f\\ v_0\end{pmatrix},$$

$$F(\eta)=\begin{pmatrix}0\\ \varepsilon\psi_\delta(x,\bar{u})\end{pmatrix}, \qquad A=-\begin{pmatrix}0 & I\\ \Delta_n & 0\end{pmatrix},$$

we may write (2.1) as the equivalent system

$$\begin{aligned}&\eta'(t)=-A\eta(t)+F(\eta(t)), \qquad 0<t<T,\\ &\eta(0)=\eta_0.\end{aligned}\tag{2.2}$$

Let H denote the Hilbert space of real-valued functions $H_0^1(D)\oplus L_2(D)$, with inner product

$$\left\langle\begin{pmatrix}u\\ v\end{pmatrix},\begin{pmatrix}w\\ z\end{pmatrix}\right\rangle=\int_D\nabla u\cdot\nabla w\,dx+\int_D vz\,dx.\tag{2.3}$$

Provided ∂D is of class C^2, A is a closed skew-adjoint operator on H with domain $\operatorname{dom}(A)\equiv[H^2(D)\cap H_0^1(D)]\oplus H^1(D)$, and generates on H the continuous one-parameter group $W(t)=e^{-tA}$. Therefore (2.2) can be reformulated as the integral equation problem

$$\eta(t)=e^{-tA}\eta_0+\int_0^t e^{-(t-s)A}F(\eta(s))\,ds,\tag{2.4}$$

which may then be solved by the contraction mapping principle.

The following theorem summarizes results proved in [10].

THEOREM 2.5. *Let ∂D be of class C^2, and for a fixed $m\geqq 1$ let η_0 be in* $\operatorname{dom}(A^m)$. *Let $\|\cdot\|$ denote the norm on H induced by* (2.3). *Suppose that for all $1\leqq j\leqq m$,*

$$\|A^jF(\eta)\|\leqq C(\|\eta\|,\cdots,\|A^{j-1}\eta\|)\|A^j\eta\|,\tag{2.6}$$

$$\|A^j(F(\eta)-F(\nu))\| \leqq C(\|\eta\|, \|\nu\|, \cdots, \|A^j\eta\|, \|A^j\nu\|)\|A^j(\eta-\nu)\| \tag{2.7}$$

for all η, ν *in* $\operatorname{dom}(A^j)$, *where the constants* C *are nondecreasing, everywhere finite functions of all their variables. Then there is a* $T>0$ *such that* (2.2) *has a unique continuously differentiable solution* $\eta(t)$ *for* $0 \leqq t < T$, *with* $\phi(t)$ *in* $\operatorname{dom}(A^m)$ *for all* $0 \leqq t < T$.

If in addition $\|\eta(t)\|$ *is bounded on any finite interval of existence of* $\eta(t)$, *then* $\eta(t)$ *exists globally in time.*

Let $\|\cdot\|_p$ denote the norm in $L_p(D)$, and let $K_1, K_2, \cdots$ denote positive constants. If $\eta \in \operatorname{dom}(A)$ has first component u, then

$$\begin{aligned}\|AF(\eta)\|^2 &= \varepsilon^2 \int_D |\nabla \psi_\delta(x, u)|^2\, dx \\ &\leqq K_1[\|\phi_\delta'(u+f)|\nabla u|\|_2^2 + \||\phi_\delta'(u+f)-\phi_\delta'(f)||\nabla f|\|_2^2] \\ &\leqq K_2[\|\nabla u\|_2^2 + \|u\|_2^2] \leqq K_3\|\nabla u\|_2^2 \leqq K_4\|A\eta\|^2,\end{aligned}$$

where use was made of the boundedness and uniform Lipschitz continuity of ϕ_δ' on $\mathbb{R}$, the boundedness of $|\nabla f|$ on $\bar{D}$, and the Poincaré inequalities

$$\|u\|_2 \leqq K\|\nabla u\|_2 \leqq K^2\|\Delta_n u\|_2 \tag{2.8}$$

valid for $u \in H^2(D) \cap H_0^1(D)$. This establishes (2.6) with $j=1$.

If η, $\nu \in \operatorname{dom}(A)$ have respective first components u, w, then by applying Hölder's inequality, the Sobolev inequality

$$\|u\|_p \leqq C_p\|\nabla u\|_2, \qquad 1 \leqq p \leqq \frac{2n}{n-2} \tag{2.9}$$

valid for $u \in H^1(D)$, and (2.8), we obtain

$$\begin{aligned}\|A(F(\eta)-F(\nu))\| &\leqq K_1[\|u-w\|_2^2 + \|\nabla(u-w)\|_2^2 + \|\nabla u\|_4^2\|u-w\|_4^2] \\ &\leqq K_2\|\Delta_n(u-w)\|_2^2 + K_3\|\Delta_n u\|_2^2\|\nabla(u-w)\|_2^2 \\ &\leqq K_4(1+\|A\eta\|^2)\|A(\eta-\nu)\|^2.\end{aligned}$$

This establishes (2.7) with $j=1$ for $1 \leqq n \leqq 4$.

From the integral equation (2.4), from the fact that $|\psi_\delta(x, u)| \leqq C|u|$ for some constant $C>0$ for all $x \in \bar{D}$ and $u \in \mathbb{R}$, and from (2.8), we obtain

$$\begin{aligned}\|\eta(t)\| &\leqq \|e^{-tA}\eta_0\| + \left\| \int_0^t e^{-(t-s)A}F(\eta(s))\, ds \right\| \\ &\leqq \|\eta_0\| + \int_0^t \|F(\eta(s))\|\, ds \\ &\leqq K_1\left[\|\eta_0\| + \int_0^t \|\eta(s)\|\, ds\right];\end{aligned}$$

hence, by Gronwall's inequality,

$$\|\eta(t)\| \leqq K_1\|\eta_0\| e^{K_1 t}$$

for all t in the existence interval for $\eta(t)$.

The above arguments and Theorem 2.5 together yield the following corollary.

COROLLARY 2.10. *Let* $D \subseteq \mathbb{R}^n$ *with* $1 \leqq n \leqq 4$ *have a* C^2 *boundary, and suppose that* $\eta_0 \in \operatorname{dom}(A) = [H^2(D) \cap H_0^1(D)] \oplus H^1(D)$. *Then for all* $0 < \delta < M$ *and for all sufficiently small* $\varepsilon > 0$, *problem* (2.2) *has a unique continuously differentiable solution* $\eta(t)$ *that is global in time and remains in* $\operatorname{dom}(A)$ *for all* $t \geqq 0$.

Four remarks are in order. Note that the proof of Corollary 2.10 does not require ϕ to be convex. Theorem 2.5 cannot be used to obtain greater regularity of solutions of (2.2) due to the lack of a Poincaré inequality of the form

$$\|\nabla^j u\|_2 \leqq C\|\nabla^{j+1} u\|_2, \qquad u \in H^{j+1}(D) \cap H_0^1(D)$$

for $j \geqq 2$. The first component $\bar{u}$ of the solution $\eta = \binom{\bar{u}}{\bar{u}_t}$ in Corollary 2.10 with $n = 2$ or 3 is continuous on $\bar{D} \times [0, \infty)$ in view of the imbedding inequality

$$\max_{x \in \bar{D}} |\bar{u}(x, t)| \leqq C\|\Delta_n \bar{u}(\cdot, t)\|_2^2$$

valid for $\bar{u}(\cdot, t) \in H^2(D) \cap H_0^1(D)$, $1 \leqq n \leqq 3$, and the continuity in time of $\bar{u}$ in the norm on $H^2(D) \cap H_0^1(D)$. Since $H^2(D)$ imbeds in $C(\bar{D})$ only for $n = 1, 2, 3$ [3, p. 30], any extension of Corollary 2.10 to cases $n \geqq 5$ would not be useful for the purposes of this paper.

Now suppose that the solution η in Corollary 2.10 with $n = 2$ or 3 begins with $\max_{x \in \bar{D}} [u_0(x)] < M - \delta$. If $\max_{x \in \bar{D}} [\bar{u}(x, t) + f(x)] < M - \delta$ for all $t \geqq 0$, then $u = \bar{u} + f$ is a global solution of problem (An). Otherwise there is a first time $T_0 > 0$ at which $\max_{x \in \bar{D}} [\bar{u}(x, T_0) + f(x)] = M - \delta$; by choosing $0 < \delta_1 < \delta$ and applying Corollary 2.10 to problem (2.2) with δ replaced by δ_1, we may extend $\bar{u} + f$ uniquely to an interval $[0, T_1)$ with $T_1 > T_0$ such that $\max_{x \in \bar{D}} [\bar{u}(x, t) + f(x)] < M - \delta_1$ for $0 \leqq t < T_1$, and $u = \bar{u} + f$ solves (An) on $\bar{D} \times [0, T_1)$. This argument establishes the following corollary.

COROLLARY 2.11. *Let $D \subseteq \mathbb{R}^n$ with $n = 2$ or 3 have a C^2 boundary, and suppose that $\binom{u_0}{v_0} \in \operatorname{dom}(A)$ with $\max_{x \in \bar{D}} u_0(x) < M$. Then for all sufficiently small $\varepsilon > 0$, the system form of problem* (An)

$$\begin{pmatrix} u \\ u_t \end{pmatrix}' = -A \begin{pmatrix} u \\ u_t \end{pmatrix} + \begin{pmatrix} 0 \\ \varepsilon\phi(u) \end{pmatrix}, \qquad 0 < t < T,$$

$$\begin{pmatrix} u \\ u_t \end{pmatrix}\Bigg|_{t=0} = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}$$

has a unique solution $\binom{u}{u_t} \in \operatorname{dom}(A)$ on some time interval $[0, T)$ that is continuously differentiable in time in the norm on H. The solution can be continued as long as $\max_{x \in \bar{D}} u(x, t) < M$.

Define $\Phi(u) = \int_0^u \phi(s)\, ds$. A solution u of (An) with the regularity properties given in Corollary 2.11 satisfies the energy equality

$$E(t) \equiv \frac{1}{2} \int_D |u_t|^2\, dx + J(u) = E(0), \tag{2.12}$$

where

$$J(u) = \frac{1}{2} \int_D |\nabla u|^2\, dx - \varepsilon \int_D \Phi(u)\, dx$$

represents the potential energy of u at time t.

By defining $j(\lambda) = J(f + \lambda u)$ for $\lambda \geqq 0$, we may examine the behavior of J along rays emanating from the stationary solution f of problem (An) in the function space $H_0^1(D)$. Note that

$$\begin{aligned} j'(0) &= \int_D \nabla f \cdot \nabla u\, dx - \varepsilon \int_D \phi(f) u\, dx \\ &= \int_D \nabla f \cdot \nabla u\, dx + \int_D (\Delta f) u\, dx = 0. \end{aligned}$$

Let $f_{\max} = \max_{x \in \bar{D}} f(x)$; then by (2.8),

$$j''(0) = \int_D |\nabla u|^2 \, dx - \varepsilon \int_D \phi'(f) u^2 \, dx$$

$$\geqq \int_D |\nabla u|^2 \, dx - \varepsilon \phi'(f_{\max}) \int_D u^2 \, dx$$

$$\geqq (1 - \varepsilon K^2 \phi'(f_{\max})) \int_D |\nabla u|^2 \, dx.$$

Suppose f satisfies $f_{\max} \to 0$ as $\varepsilon \to 0^+$. (Equilibrium solutions with this property do exist; see [5].) Then for all sufficiently small ε we have $1 - \varepsilon K^2 \phi'(f_{\max}) > 0$, and hence $j''(0) > 0$; i.e., J is convex along rays emanating from f. This is a necessary condition for f being a local minimum of J in $H_0^1(D)$.

Nevertheless, as the following results show, there exists no potential well about any stationary solution of (An) for any $\varepsilon > 0$ and any $n \geqq 2$.

LEMMA 2.13. *Let $\varepsilon > 0$ be fixed, and let f be an equilibrium solution of* (An), *$n \geqq 2$. Choose $x_0 \in D$ such that $f(x_0) = f_{\max}$. Then for any $\delta > 0$ we may find a ball $B_M \subseteq D$ with center x_0, and functions $\{w_\lambda : f_{\max} < \lambda \leqq M\} \subseteq C_0^\infty(D)$, such that $w_\lambda = \lambda - f$ on B_M and $\int_D |\nabla w_\lambda|^2 \, dx < \delta$ for all $f_{\max} < \lambda \leqq M$.*

Proof. Let $B(x_0, a)$, $B(x_0, b)$ denote concentric open balls in $\mathbb{R}^n$ with center x_0 and respective radii $0 < a < b$. By Friedman [3, p. 9] there exists a $\zeta \in C^\infty(\mathbb{R}^n)$ such that $\zeta = 1$ in $B(x_0, a)$, $0 \leqq \zeta \leqq 1$ in $B(x_0, b) - B(x_0, a)$, and $\zeta = 0$ outside $B(x_0, b)$. The function ζ will be called a C^∞ cutoff from $B(x_0, a)$ to $B(x_0, b)$. Note that for $0 < \rho < 1$, $\zeta_\rho(x) \equiv \zeta(x/\rho)$ is a C^∞ cutoff from $B(x_0, a\rho)$ to $B(x_0, b\rho)$ satisfying $|\nabla \zeta_\rho(x)| \leqq K_0/\rho$ for all $x \in \mathbb{R}^n$, where $K_0 > 0$ depends on a, b, but is independent of ρ. □

The proof for $n \geqq 3$ proceeds as follows. Choose $0 < a < b$, and for each $f_{\max} \leqq \lambda \leqq M$ and each $\rho > 0$ define

$$w_{\lambda,\rho} = \zeta_\rho(\lambda - f).$$

Then for all ρ sufficiently small $w_{\lambda,\rho} \in C_0^\infty(D)$, and $w_{\lambda,\rho} = \lambda - f$ on $B(x_0, a\rho)$. Now f is nonnegative on $\bar{D}$ by the maximum principle, so for $f_{\max} \leqq \lambda \leqq M$

$$\int_D |\nabla w_{\lambda,\rho}|^2 \, dx = 2 \int_{B(x_0, b\rho)} [|\nabla \zeta_\rho|^2 |\lambda - f|^2 + |\zeta_\rho|^2 |\nabla f|^2] \, dx$$

$$\leqq 2 V_n (K_0^2 M^2 + \rho^2 \max_{\bar{D}} |\nabla f|^2) \rho^{n-2},$$

where V_n denotes the volume of the unit ball in $\mathbb{R}^n$. By taking $\rho = \rho_0 > 0$ sufficiently small, we may arrange that

$$\int_D |\nabla w_{\lambda,\rho_0}|^2 \, dx < \delta \quad \text{for all } f_{\max} \leqq \lambda \leqq M.$$

We may then take B_M to be $B(x_0, a\rho_0)$, and w_λ to be w_{λ,ρ_0} for $f_{\max} \leqq \lambda \leqq M$.

When $n = 2$, choose fixed $b > a > 0$ so small that $B(x_0, b) \subseteq D$. Let $r = |x - x_0|$ denote the distance from x_0 to x in $\mathbb{R}^2$. For $f_{\max} < \lambda \leqq M$, choose $B \equiv \lambda - f_{\max} > A > 0$, and choose $0 < 2\rho < a$. For $\alpha < 0$ let

$$C = \frac{B - A}{\rho^\alpha - b^\alpha}, \qquad D = \frac{A\rho^\alpha - Bb^\alpha}{\rho^\alpha - b^\alpha},$$

so that $C\rho^\alpha + D = B$ and $Cb^\alpha + D = A$. Let ζ denote a C^∞ cutoff from $B(x_0, 1)$ to $B(x_0, 2)$, and define $\zeta_\rho(r) = \zeta(r/\rho)$. Let $\zeta_{a,b}$ be a C^∞ cutoff from $B(x_0, a)$ to $B(x_0, b)$. For $f_{\max} < \lambda \leqq M$ define

$$w_{\lambda,\rho}(x) = \zeta_{a,b}(r)[\zeta_\rho(r)(\lambda - f(x) - Cr^\alpha - D) + Cr^\alpha + D].$$

Thus setting

$$I_1 = \int_{B(x_0,2\rho)} |\nabla f|^2 \, dx,$$

$$I_2 = \int_{B(x_0,2\rho)-B(x_0,\rho)} |\nabla \zeta_\rho|^2 |\lambda - f - Cr^\alpha - D|^2 \, dx,$$

$$I_3 = \int_\rho^b |\nabla(Cr^\alpha)|^2 r \, dr,$$

$$I_4 = \int_a^b |\nabla \zeta_{a,b}|^2 |Cr^\alpha + D|^2 r \, dr,$$

we may obtain that

$$\int_D |\nabla w_{\lambda,\rho}|^2 \, dx \leqq K_1 \sum_{j=1}^4 I_j$$

for an absolute constant $K_1 > 0$.

Since $|\nabla f|$ is bounded on $B(x_0, b)$, we may choose a $\rho_0 > 0$ independent of A, α, and λ such that $I_4 < \delta/(4K_1)$ for all $0 < \rho \leqq \rho_0$.

Using the facts that $Cr^\alpha + D$ is positive and decreasing for $0 < r \leqq b$, and that $|\nabla \zeta_\rho| \leqq K_0/\rho$ for some absolute constant $K_0 > 0$, we may show that

$$I_2 \leqq 6\pi K_0^2 [C^2(1-2^\alpha)^2 \rho^{2\alpha} + \max_{x \in \overline{B(x_0,2\rho)}} |f(x) - f_{\max}|^2],$$

$$I_3 = \frac{\alpha C^2}{2}(b^{2\alpha} - \rho^{2\alpha}), \qquad I_4 \leqq K_2[A^2 + C^2(a^\alpha - b^\alpha)^2],$$

for some positive absolute constant K_2. It is a simple matter to show that the expressions $C^2(1-2^\alpha)^2\rho^{2\alpha}$, $(\alpha C^2/2)(b^{2\alpha} - \rho^{2\alpha})$, $C^2(a^\alpha - b^\alpha)^2$ can be made arbitrarily small independent of λ, $f_{\max} < \lambda \leqq M$, by taking $|\alpha| > 0$, $\rho > 0$ sufficiently small. Hence by choosing $A = A_0$, $\rho = \rho_1 < \rho_0$, $\alpha = \alpha_0$ all sufficiently close to zero, we may arrange that $I_j < \delta/4K_1$ for $1 \leqq j \leqq 4$ for all $f_{\max} < \lambda \leqq M$. Set $B_M = B(x_0, \rho_1)$, and $w_\lambda = w_{\lambda,\rho_1}$ with $A = A_0$, $\alpha = \alpha_0$. Then $\int_D |\nabla w_\lambda|^2 \, dx < \delta$ for all $f_{\max} < \lambda \leqq M$. □

In particular, for any $\varepsilon > 0$, $n \geqq 2$, there are functions $v_M = w_M + f$ with essential supremum equal to M in any neighborhood of f in $H_0^1(D)$.

LEMMA 2.14. *Let $y(t)$ denote the solution of the ordinary initial boundary value problem*

$$\begin{aligned} &\ddot{y} = \varepsilon \phi(y) \qquad t > 0, \\ &y(0) = y_0, \qquad \dot{y}(0) = 0, \end{aligned} \tag{2.15}$$

where $y_0 < M$. Then there is a finite $T_q > 0$ such that $\lim_{t \to T_q^-} y(t) = M$, i.e., y quenches in finite time. As $y_0 \to M^-$, $T_q \to 0^+$.

Proof. The uniform Lipschitz continuity of ϕ on intervals of the form $(-\infty, M - \delta]$, $\delta > 0$, guarantees that (2.15) has a unique local C^2 solution $y(t)$ that can be continued as long as $y(t) < M$. On the existence interval $[0, T_q)$ for y, $\ddot{y} > 0$ and hence $\dot{y}(t)$ is

strictly increasing in t. Since $\dot{y}(0)=0$, $\dot{y}(t)>0$ and hence $y(t)>y_0$ for $0<t<T_q$. Since ϕ is strictly increasing,

$$\ddot{y}(t)>\varepsilon\phi(y_0),$$

so that

$$y(t)-y_0>\varepsilon\phi(y_0)t^2,$$

for $0\leqq t<T_q$. Hence

$$\sqrt{(M-y_0)/\varepsilon\phi(y_0)}>T_q.$$

Clearly, $T_q\to 0^+$ as $y_0\to M^-$. □

THEOREM 2.16. *Let ε, δ, T_0 be any fixed positive numbers, and let k be any nonnegative integer. Let $f\in C^k(\bar{D})$ be an equilibrium solution of* (An) *with $n\geqq 2$. Then there exists $u_0\in C^k(\bar{D})$ with $u_0=0$ on ∂D, $\max_{\bar{D}} u_0<M$, and $\int_D|\nabla(u_0-f)|^2\,dx<\delta$, such that the solution u of problem* (An) *with $v_0=0$ quenches in finite time $T\leqq T_0$.*

Proof. Let $w_\lambda\in C_0^\infty(\bar{D})$ be the functions satisfying $w_\lambda=\lambda-f$ on B_M, $\int_D|\nabla w_\lambda|^2\,dx<\delta$ for all $f_{\max}<\lambda\leqq M$, whose existence is guaranteed by Lemma 2.13. Let $\rho>0$ denote the radius of B_M. By Lemma 2.14, we may choose $y_0<M$ so close to M that the solution $y=y(t,y_0)$ of (2.15) quenches in time $T_q\leqq\min\{T_0,\rho\}$.

Define $u_0=w_{y_0}+f$; then $u_0\in C^k(\bar{D})$, $u_0=0$ on ∂D, $u_0=y_0$ on B_M, and $\int_D|\nabla(u_0-f)|^2\,dx<\delta$. If u denotes the solution of problem (An) with this u_0 and with $v_0=0$, we have

$$u(x,t)=y(t,y_0)$$

for all (x,t) in the retrograde characteristic cone with vertex (x_0,ρ) and base $B_M\times\{0\}$. Hence u must quench in time $T\leqq T_q$. □

The idea of comparing solutions inside retrograde characteristic cones in the half-space $t\geqq 0$ was used by Keller [6] to show pointwise blowup in finite time of solutions of $u_{tt}=c^2\Delta_n u+f(u)$ for certain $f\in C^2(\mathbb{R})$.

Suppose u is a continuous solution of (An) with sufficient regularity to satisfy energy equality (2.12) (or the inequality $E(t)\leqq E(0)$) for all t in its existence interval. Define

$$u^+_{\max}=u^+_{\max}(t)\equiv\max_{x\in\bar{D}}(u(x,t),0),\qquad \gamma=\gamma(t)\equiv(u^+_{\max})^{-2}\int_D|\nabla u|^2\,dx.$$

Then

$$\gamma\leqq 2(u^+_{\max})^{-2}[E(0)+\varepsilon\mu(D)\Phi(u^+_{\max})]\equiv g(u^+_{\max}), \tag{2.17}$$

where $\mu(D)$ denotes the n-dimensional Lebesgue measure of D.

When $E(0)\geqq 0$, $g(s)$ achieves a positive absolute minimum $g_m=g(s_0)$ on the interval $(0,M]$. Note that $g(M/2)$, and hence g_m itself, can be made arbitrarily small by taking both $\varepsilon>0$ and $E(0)\geqq 0$ sufficiently small.

If u satisfies an a priori inequality of the form

$$\gamma(t)\geqq g_m,\qquad t\geqq 0 \tag{2.18}$$

and begins in the region R depicted in Fig. 1 (i.e., with $u^+_{0,\max}<s_0$), then $u^+_{\max}$ remains bounded away from M for all time, and u will be a global solution of (An).

When $n=1$ we have $\gamma(t)\geqq 4$ for all $t\geqq 0$ by (1.1), and these observations underlie the proof of the result (b) of [2, §1].

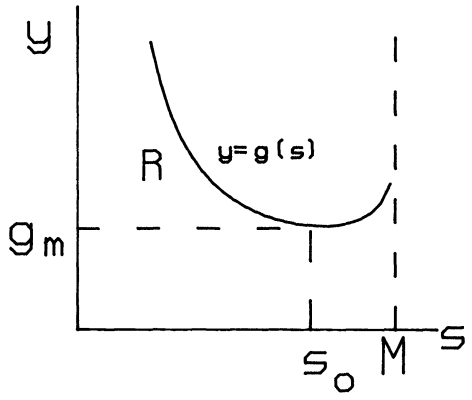


FIG. 1. *An invariant region for solutions of* (An) *satisfying* (2.18).

3. Numerical results. A simple explicit finite-difference scheme was used to approximate solutions $u = u(r, t)$ of

(Rn)

$$u_{tt} = u_{rr} + \frac{n-1}{r} u_r + \varepsilon\phi(u), \quad 0 < r < 1, \quad 0 < t < T, \tag{3.1}$$

$$u_r(0, t) = u(1, t) = 0, \qquad 0 < t < T, \tag{3.2}$$

$$u(r, 0) = u_0(r), \quad u_t(r, 0) = v_0(r), \quad 0 < r < 1, \tag{3.3}$$

which are radial solutions of (An) when $r = |x|$ and D is the unit ball centered at the origin in $\mathbb{R}^n$.

The difference scheme used is adapted from John [4, pp. 172–174]. Divide the interval $[0, 1]$ into N subintervals of equal length $h = 1/N$, and let k denote the stepsize in time, with

$$\lambda \equiv \frac{k}{h} \leqq 1. \tag{3.4}$$

For $1 \leqq i \leqq N+1$, $j \geqq 0$, $w = w(r, t)$ define

$$w_{ij} = w((i-1)h, jk).$$

Let δ_t denote the divided difference operator

$$\delta_t w_{ij} = \frac{1}{k}\left[w_{i,j+1} - \frac{1}{2}(w_{i+1,j} + w_{i-1,j}) \right],$$

with space averaging in the lower step; and let δ_r denote the central divided difference operator

$$\delta_r w_{ij} = \frac{1}{2h}[w_{i+1,j} - w_{i-1,j}].$$

For $2 \leqq i \leqq N-1$ and $j \geqq 0$, (3.1) was replaced by the difference equation

$$\delta_t^2 w_{ij} = \delta_r^2 w_{ij} + \frac{n-1}{(i-1)h} \delta_r w_{ij} + \varepsilon\phi(w_{ij}). \tag{3.5}$$

Values $w_{0,j}, w_{-1,j}, \cdots$ were interpreted by extending w as an even function of r through $r = 0$. For $i = 1$, (3.1) was replaced by

$$\delta_t^2 w_{1j} = n\delta_r^2 w_{1j} + \varepsilon\phi(w_{1j}). \tag{3.6}$$

For $i = N$, backward differences in r were used, and space averaging was abandoned wherever necessary to avoid going past $r = 1$. A Taylor series approximation was used to obtain the values w_{i1}.

The difference scheme is stable, consistent, and convergent when applied to the pure initial boundary value problem obtained by linearizing (3.1) about a stationary solution f. Grave difficulties are encountered, however, in attempts to prove consistency and convergence for (3.4)-(3.6), due to the boundary conditions and the presence of the nonlinearity ϕ. In particular, we are unable to derive useful upper bounds for higher difference quotients of w. This is analogous to the difficulties encountered with the abstract approach to the differential problem in § 2. Therefore for the numerical tests the following checks and safeguards were implemented.

(a) The Courant-Friedrichs-Lewy condition (3.4), a necessary condition for stability, was ensured to be satisfied by taking $\lambda = \frac{1}{4}$ in all tests;

(b) Stationary solutions f of (Rn) were approximated by the shooting method using the classical fourth-order Runge-Kutta method. The difference scheme (3.4)-(3.6) was then applied with $u_0 = f$, $v_0 = 0$ as a check of the computer code. Since the

TABLE 1
Values of ε_n, ε_* *for* $\phi(u) = (1-u)^{-1}$.

n	ε_n	ε_*
1	0.341	0.383
2	1.017	1.309
3	1.520	2.139
7	2.563	6.000

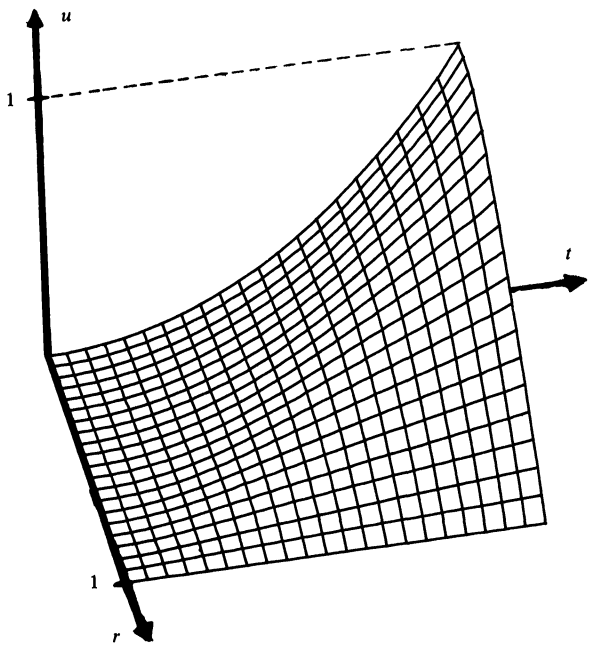


FIG. 2. *A solution of* (3.5), (3.6) *with* $\varepsilon > \varepsilon_n$.

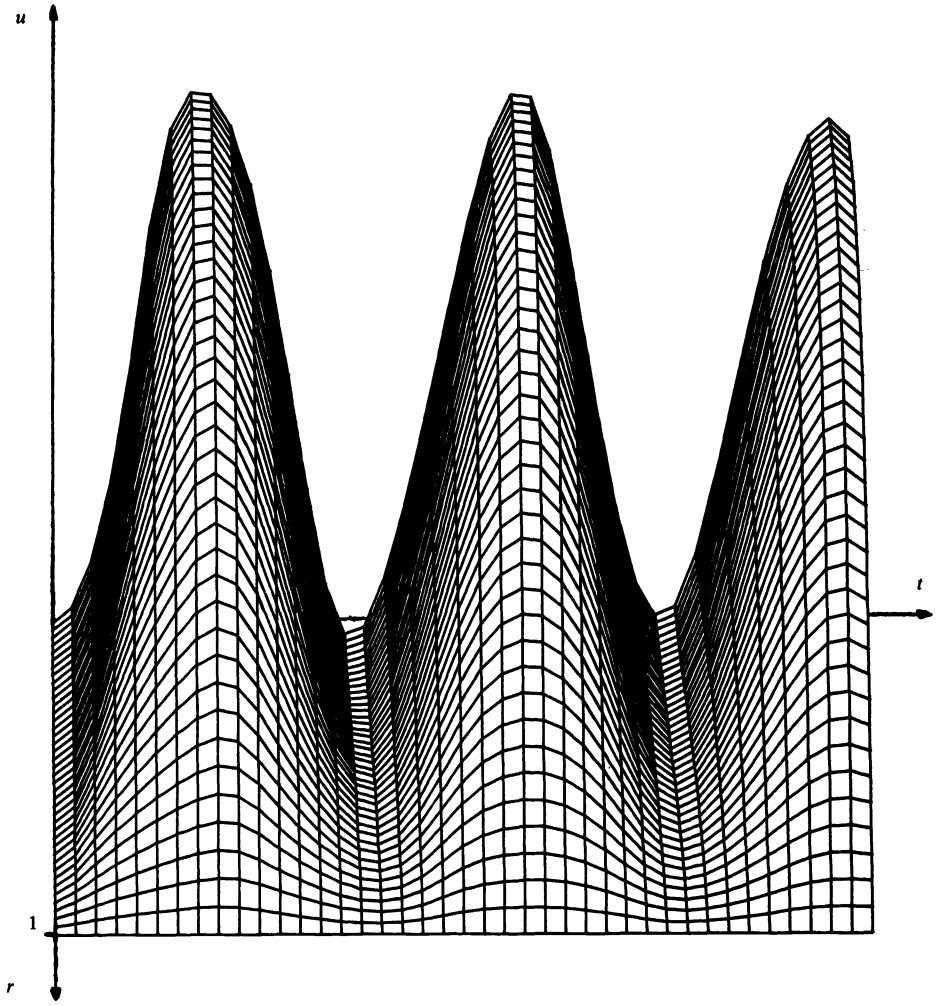


FIG. 3. *A solution of* (3.5), (3.6) *with* $\varepsilon < \varepsilon_n$.

approximate stationary solutions are not exact, these checks (as well as checks with $v_0 = 0$, $u_0 =$ small perturbation of f) served as empirical evidence of the scheme's stability;

(c) The convergence of the scheme (3.5), (3.6) was checked empirically for several examples by letting h, $k \to 0$ while keeping $\lambda = \frac{1}{4}$; and

(d) The difference scheme satisfies an energy inequality related to (2.12). Care was taken so that the total energy of the difference approximation remained nearly constant.

Double-precision arithmetic was used for all computations. The experiments were performed on a National Advanced Systems AS/9160 computer with MVS/SP operating system.

To isolate the effects of the term $\varepsilon\phi(u)$, solutions of (3.5), (3.6) with $u_0 = v_0 = 0$ were computed in dimensions $n = 2$, 3, and 7. The behavior of such solutions agrees qualitatively with behavior reported in [2] for solutions in the case $n = 1$. In particular, in each dimension n considered there appears to be an $\varepsilon_n > 0$ such that solutions

quench in finite time when $\varepsilon > \varepsilon_n$ and do not quench (even in infinite time) when $\varepsilon < \varepsilon_n$. For $\varepsilon < \varepsilon_n$ the solution displays a sequence of relative maxima that occur along the line $r = 0$, $t > 0$; the first such relative maximum appears to be an absolute maximum which approaches 1 from below as ε approaches ε_n from below.

Table 1 lists values of ε_n obtained when $\phi(u) = (1-u)^{-1}$; the value of ε_1 is taken from [2].

Figures 2 and 3 contrast the behavior of a solution w of (3.5), (3.6) for values of ε, respectively, greater than and less than ε_n. Figures 2 and 3 were generated using $\phi(u) = (1-u)^{-1}$, $n = 2$, and $h = 1/200$. In Fig. 2, $\varepsilon = 1.5 > \varepsilon_2$, and the solution quenches in time $T = 1.01$, while in Fig. 3, $\varepsilon = 0.9 < \varepsilon_2$, and the solution is displayed for $0 \leqq t \leqq 8$.

The stability of the solution f of the stationary problem

$$f(r) + \frac{n-1}{r} f'(r) + \varepsilon\phi(f(r)) = 0, \qquad 0 < r < 1, \tag{3.7}$$

$$f'(0) = f(1) = 0, \tag{3.8}$$

satisfying $f_{\max} \to 0^+$ as $\varepsilon \to 0^+$, was also investigated. Care must be taken in computing f, since solutions of (3.7), (3.8), are not always unique. Indeed, in [5], Joseph and Lundgren show that for $\phi(u) = (1+\alpha u)^{\beta}$ with α, $\beta < 0$, $\tau = 1/(\beta - 1)$, $\bar{\varepsilon} = (\tau/\alpha)(n-2-\tau)$, and $f(\beta) = 2\beta\tau + 2(2\beta\tau)^{1/2}$:

(a) There is an $\varepsilon_* > 0$ such that positive solutions of (3.7), (3.8) do not exist when $\varepsilon > \varepsilon_*$;

(b) For $3 \leqq n < 2 + f(\beta)$ and $\varepsilon_* > \bar{\varepsilon}$, there is a large but finite number of positive solutions when $\varepsilon < \bar{\varepsilon}$ is close to $\bar{\varepsilon}$, and a countably infinite number of solutions when $\varepsilon = \bar{\varepsilon}$; and

(c) For $n \geqq 2 + f(\beta)$ and $\bar{\varepsilon} = \varepsilon_*$, there is exactly one positive solution when $\varepsilon < \varepsilon_*$. See Table 1 for values of ε_* when $\phi(u) = (1-u)^{-1}$ and $n = 1, 2, 3, 7$.

Bifurcation diagrams plotting ε as a function of $f_{\max}$ were generated using the procedure described in [5]. We then checked that the shooting method converged to the stationary solution with smallest $f_{\max}$. Figure 4 contains bifurcation diagrams for $n = 2, 3, 9$ when $\alpha = -1$, $\beta = -3$.

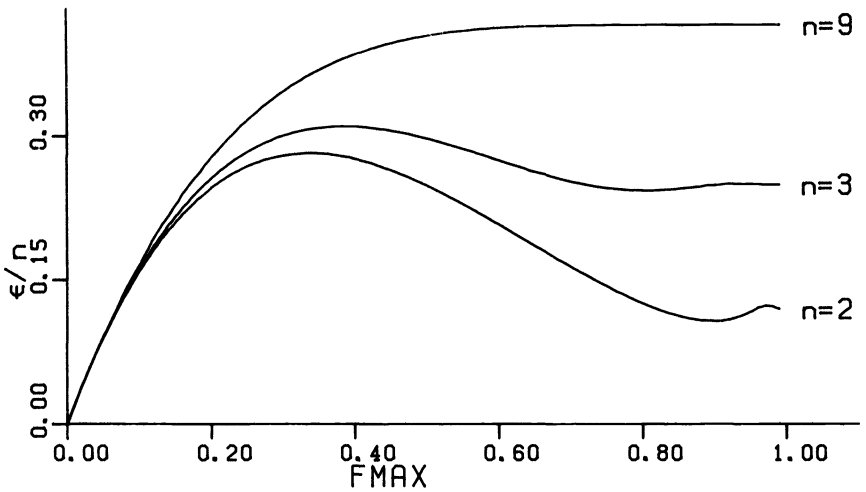


FIG. 4. *Bifurcation diagram for positive solutions of* (3.7), (3.8) *when* $\phi(u) = (1-u)^{-3}$.

Perturbations p of f used as initial data u_0 were of the form $p = vf$ for $1 < v < 1/f_{\max}$ or of the form

$$p(r) = \begin{cases} v, & 0 \leqq r \leqq r_0, \\ f(r), & r_1 \leqq r \leqq 1, \end{cases}$$

where $0 < r_0 < r_1 < 1$, $f_{\max} < v < 1$, and p is defined on $r_0 \leqq r \leqq r_1$ to be strictly decreasing and C^2 on $[0, 1]$. Numerical experiments with initial data $u_0 = p$, $v_0 = 0$ indicate that whenever p is sufficiently close to f in sup-norm, the a priori inequality (2.18) is

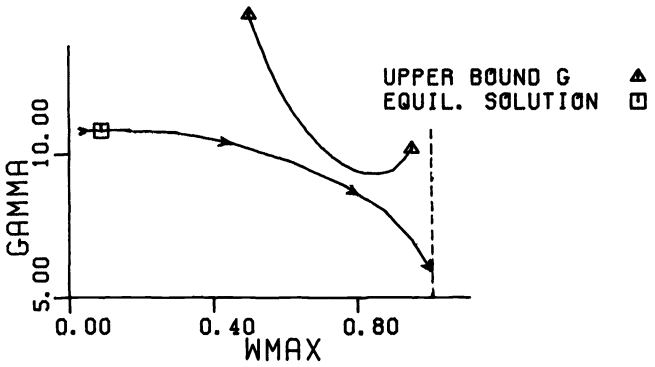


FIG. 5. *Quenching trajectory for γ when $\lambda = 0.5$, $E(0) = 2.0$.*

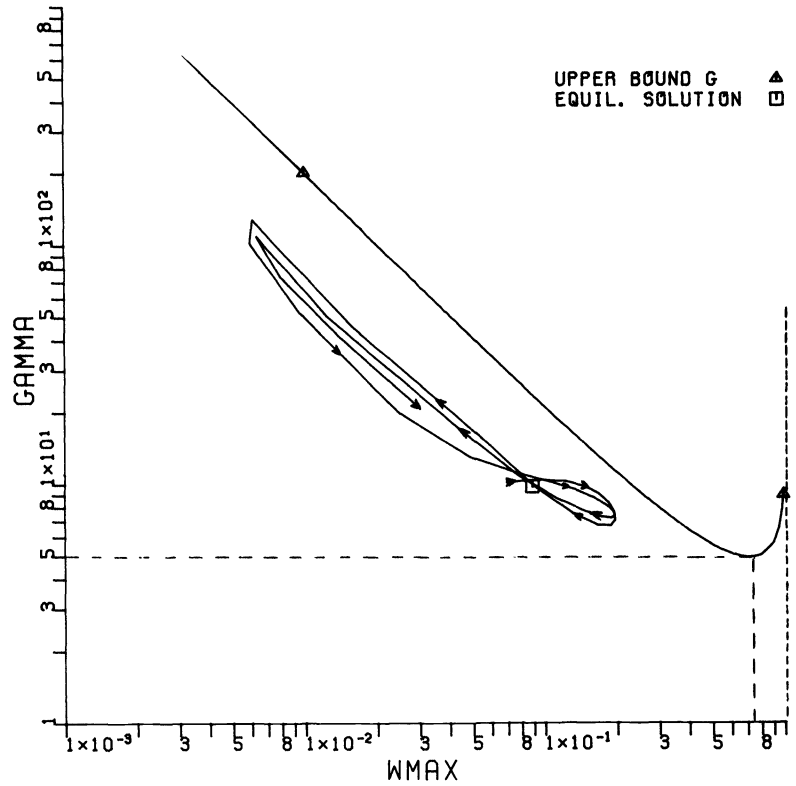


FIG. 6. *Nonquenching trajectory for γ when $\nu = 0.5$, $E(0) = 0$.*

satisfied for all time by solutions of (3.5)-(3.8). Figures 5 and 6 contrast the trajectories of $\gamma(t)$ in quenching and nonquenching cases. Figures 5 and 6 were generated with $n=3$, $\phi(u)=(1-u)^{-1}$, $u_0=vf$, $v_0(r)=\mu(1-r^2)$, where μ was chosen so that $E(0)\geqq 0$, $\varepsilon=0.5$, $h=1/100$.

Acknowledgment. The author thanks Howard A. Levine for many helpful discussions.

REFERENCES

[1] A. ACKER AND W. WALTER, *On the global existence of solutions of parabolic differential equations with a singular nonlinear term*, Nonlinear Anal., 2 (1978), pp. 499-505.

[2] P. H. CHANG AND H. A. LEVINE, *The quenching of solutions of semilinear hyperbolic equations*, SIAM J. Math. Anal., 12 (1982), pp. 893-903.

[3] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart, and Winston, New York, 1969.

[4] F. JOHN, *Partial Differential Equations*, Fourth edition, Appl. Math. Sci. 1, Springer-Verlag, Berlin, New York, 1982.

[5] D. D. JOSEPH AND T. S. LUNDGREN, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Rational Mech. Anal., 49 (1972), pp. 241-269.

[6] J. B. KELLER, *On solutions of nonlinear wave equations*, Comm. Pure Appl. Math., 10 (1957), pp. 523-530.

[7] H. A. LEVINE, *The phenomenon of quenching: A survey*, in Proc. VIth International Conference on Trends in the Theory and Practice of Non-linear Analysis, V. Lakshmikantham, ed., Elsevier North-Holland, New York, 1985.

[8] ———, *The quenching of solutions of linear parabolic and hyperbolic equations with nonlinear boundary conditions*, SIAM J. Math. Anal., 14 (1983), pp. 1139-1153.

[9] H. A. LEVINE AND M. W. SMILEY, *The quenching of solutions of linear parabolic and hyperbolic equations with nonlinear boundary conditions*, J. Math. Anal. Appl. 103 (1984), pp. 409-427.

[10] M. REED, *Abstract Nonlinear Wave Equations*, Lecture Notes in Mathematics 507, Springer-Verlag, Berlin, New York, 1976.

[11] D. H. SATTINGER, *On global solution of nonlinear hyperbolic equations*, Arch. Rational Mech. Anal., 30 (1968), pp. 148-172.

[12] S. D. ZAIDMAN, *Abstract Differential Equations*, Research Notes in Math. 36, Pitman, San Francisco, 1979.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1095-1108, September 1989

005

ISOPERIMETRIC INEQUALITIES FOR THE STEFAN PROBLEM*

B. GUSTAFSSON† AND J. MOSSINO‡

Abstract. The weak solution (θ, h) of the Stefan problem in some annular domain $\omega \times (0, T)$ is compared with the weak solution (Θ, H) of the "symmetrized" problem, in $\Omega \times (0, T)$, where Ω is a symmetrical annulus having the same measure as ω. For the one-phase Stefan problem—$\theta \geqq 0$, $h \in (-\alpha, 0)$ when $\theta = 0$—it is shown in particular that the "volume of ice" (meas $\{h(t) = -\alpha\}$) remains greatest in spherical symmetry (with initial data decreasing along the radii).

Résumé. Cet article compare la solution faible (θ, h) du problème de Stefan dans un domaine $\omega \times (0, T)$ (où ω est une couronne) avec celle du problème "symétrisé" (Θ, H) dans $\Omega \times (0, T)$, où Ω est une couronne symétrique de même mesure que ω. Pour le problème de Stefan à une phase—$\theta \geqq 0$, $h \in (-\alpha, 0)$ là où $\theta = 0$—on voit en particulier que le "volume de glace" (mes $\{h(t) = -\alpha\}$) est maximum en symétrie de révolution (avec donnée initiale décroissante le long du rayon).

Key words. one-phase Stefan problem, two-phase Stefan problem, solid phase-liquid phase, regularized problem, symmetrized problem, decreasing rearrangement, equimeasurable functions, isoperimetric inequalities

AMS(MOS) subject classifications. 35K55, 35B05

1. Introduction.

We consider the Stefan problem in its simplest form and in an annular space geometry: find a pair (θ, h) of functions defined in $q = \omega \times (0, T)$ such that, in some weak sense,

$$\begin{aligned} &\frac{\partial h}{\partial t} - \Delta\theta = 0 \quad \text{in } q, \\ &\theta = g \quad \text{on } \sigma = \partial\omega \times (0, T), \\ &h|_{t=0} = h_0, \\ &h \in a(\theta) \quad \text{a.e. in } q. \end{aligned} \tag{1.1}$$

Here we have the following:

- $\omega = \omega_0 \backslash \bar{\omega}_1$, where ω_0, ω_1 are bounded domains in $\mathbb{R}^N$ $(N \geqq 2)$ with smooth boundaries $\gamma_0 = \partial\omega_0$ and $\gamma_1 = \partial\omega_1$, and satisfying $\bar{\omega}_1 \subset \omega_0$.
- g is constant on each of $\sigma_0 = \gamma_0 \times (0, T)$ and $\sigma_1 = \gamma_1 \times (0, T)$, let us say

$$g = \begin{cases} 0 & \text{on } \sigma_0, \\ 1 & \text{on } \sigma_1. \end{cases}$$

- a is a strictly monotone graph in $\mathbb{R}^2$ (regarded as a map from $\mathbb{R}$ into subsets of $\mathbb{R}$). The typical form of a for the Stefan problem is

$$a(\theta) = \begin{cases} \alpha_0(\theta - \lambda) - \alpha & \text{for } \theta < \lambda, \\ [-\alpha, 0] & \text{for } \theta = \lambda, \\ \alpha_1(\theta - \lambda) & \text{for } \theta > \lambda, \end{cases} \tag{1.2}$$

where α, α_0, α_1 are positive constants, $\lambda \in [0, 1]$. However, our main results are valid for an arbitrary maximal monotone graph a such that $a([0, 1])$ is bounded, and such that the inverse graph $b = a^{-1}$ is a Lipschitz continuous function on $a([0, 1])$.

* Received by the editors April 15, 1988; accepted for publication (in revised form) October 10, 1988. This work was partly supported by the Swedish Natural Science Research Council (NFR) under grants R-RA 8793-100 and U-FR 8793-101.

† Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm 70, Sweden.

‡ Departement de Mathématiques, Centre National de la Recherche Scientifique, Université Paris-Sud, 91405 Orsay Cedex, France.

• $h_0 \in L^\infty(\omega)$ and satisfies an extra condition (see (1.6), (1.7) below), which essentially means that $\theta_0 = b(h_0)$ belongs to $H^1(\omega)$ and satisfies $0 \leqq \theta_0 \leqq 1$.

The physical interpretation of (1.1) when a is of the form (1.2) is that θ is the temperature and h the enthalpy of some matter that undergoes a phase change (solid-liquid) at temperature λ. The number α is proportional to the latent heat for the phase change, and α_0 and α_1 are proportional to the heat capacities of the solid and liquid states, respectively (α, α_0, α_1 are also inversely proportional to the thermal conductivity coefficients). With more general a (often single-valued), there are many other interpretations of (1.1) (e.g., porous medium equation).

Our boundary and initial data, g and h_0 above, are such that the solution (θ, h) of (1.1), by the maximum principle, will satisfy $0 \leqq \theta \leqq 1$ in all q. In the case of (1.2) with $\lambda = 0$, the temperature in the solid phase (the latter generally defined as the region where $h \leqq -\alpha$ ($h = -\alpha$ in this case)) therefore must be constantly equal to zero. Similarly, in the case $\lambda = 1$, the temperature in the liquid phase $\{h \geqq 0\}$ ($h = 0$ in this case) is constantly equal to 1. Thus, for these extreme cases, in practice we have a one-phase Stefan problem, while for $0 < \lambda < 1$, the problem really is a two-phase problem.

One standard way of making (1.1) precise is to say that (θ, h) is a weak solution of (1.1) if

$$\theta \in L^\infty(q), \quad h \in L^\infty(q), \quad h \in a(\theta) \quad \text{a.e. in } q,$$

$$\int\int_q \left(h \frac{\partial \varphi}{\partial t} + \theta \Delta \varphi \right) dx\, dt = \int_0^T \int_\gamma g \frac{\partial \varphi}{\partial \nu}\, d\gamma\, dt - \int_\omega h_0(x) \varphi(x, 0)\, dx \tag{1.3}$$

for every "test function" $\varphi \in \mathscr{C}^1(\bar{q})$ satisfying $(\partial^2 \varphi / \partial x_i\, \partial x_j) \in \mathscr{C}(\bar{q})$ and $\varphi = 0$ on $\sigma \cup (\omega \times \{T\})$ (see, e.g., [3] or [4]).

Existence and uniqueness of weak solutions can be proved in several different ways. One method, developed by Oleinik [9] (in one space dimension) and Friedman [3] (see also [4]) is to obtain the weak solution as a limit as $\varepsilon \to 0$ ($\varepsilon > 0$) of the classical solutions $(\theta_\varepsilon, h_\varepsilon)$ of some regularized problems

$$\begin{aligned} &\frac{\partial h_\varepsilon}{\partial t} - \Delta \theta_\varepsilon = 0 \quad \text{in } q, \\ &\theta_\varepsilon = g \quad \text{on } \sigma, \\ &h_{\varepsilon|_{t=0}} = h_{\varepsilon_0}, \\ &h_\varepsilon = a_\varepsilon(\theta_\varepsilon) \quad \text{in } q. \end{aligned} \tag{$1.1)_\varepsilon$}$$

Here a_ε (from $[0, 1]$ to $a([0, 1])$) are single-valued smooth functions with

$$a_\varepsilon' \geqq \delta > 0 \qquad (\delta \text{ independent of } \varepsilon) \tag{1.4}$$

such that $a_\varepsilon \to a$ as $\varepsilon \to 0$, in the sense that

$$b_\varepsilon = a_\varepsilon^{-1} \text{ (from } a([0, 1]) \text{ to } [0, 1]) \text{ converges uniformly to } b = a^{-1}. \tag{1.5}$$

Moreover, h_{ε_0} are smooth functions such that

$$h_{\varepsilon_0} \to h_0 \quad \text{in } L^1(\omega), \tag{1.6}$$

and that, in terms of $\theta_{\varepsilon_0} = b_\varepsilon(h_{\varepsilon_0})$,

$$0 \leqq \theta_{\varepsilon_0} \leqq 1 \quad \text{in } \omega, \qquad \theta_{\varepsilon_0}|_{\partial\omega} = g,$$

$$\int_\omega |\nabla \theta_{\varepsilon_0}|^2\, dx \text{ is bounded independently of } \varepsilon \quad \text{as } \varepsilon \to 0. \tag{1.7}$$

In this paper, our aim is to give isoperimetric inequalities for the Stefan problem (1.1). They are not standard, since the problem is multivalued, and the domain is doubly connected. The principal one of them, (2.1) below (from which some other inequalities follow as corollaries), is obtained by passing to the limit in the corresponding inequality $(2.1)_\varepsilon$ for $(1.1)_\varepsilon$. The proof of $(2.1)_\varepsilon$ is a direct parabolic one that relies on the techniques developed in [8] for linear parabolic problems. Some of our corollaries confirm the intuitive idea that, among all domains $\omega = \omega_0 \backslash \bar{\omega}_1$, with ω_0, ω_1, of given measures (volumes), and all equimeasurable initial data h_0, the solid "melts slowest" in the symmetrized domain, with symmetrized h_0 ($H_0 = \mathbf{H}_0 = \mathbf{h}_0$). These results were announced in a previous note [5].

One drawback of our method is that it seems to require constant boundary values for θ (on each component of the boundary), and initial data guaranteeing that $0 \leqq \theta \leqq 1$ holds throughout q (where 0 and 1 are the constant boundary values for θ).

Isoperimetric inequalities for a problem similar to $(1.1)_\varepsilon$ have been obtained earlier by Vazquez [10], in the case $\omega = \mathbb{R}^N$, by an elliptic approach (and using semigroup theory). This approach does not seem to work in our geometry.

2. Statements of results. Our isoperimetric inequalities are inequalities between certain quantities for the problems (1.1) and $(1.1)_\varepsilon$, and the corresponding quantities for certain "symmetrized problems" ($\widetilde{(1.1)}$ and $\widetilde{(1.1)}_\varepsilon$ below). Before describing these symmetrized problems, we must introduce some general notation:

- σ_N denotes the volume of the unit ball in $\mathbb{R}^N$.
- $|E|$ denotes the volume (N-dimensional Lebesgue measure) of a measurable set E in $\mathbb{R}^N$ (also: $|x| = (x_1^2 + \cdots + x_N^2)^{1/2}$ if $x \in \mathbb{R}^N$).
- $f_+ = \max\{f, 0\}$, $f_- = \max\{-f, 0\}$.
- With $\omega = \omega_0 \backslash \bar{\omega}_1$ as in § 1, Ω_j ($j = 0, 1$) denotes the open balls in $\mathbb{R}^N$ centered at the origin and having the same volumes as ω_j. Thus $\bar{\Omega}_1 \subset \Omega_0$. We also set $\Omega = \Omega_0 \backslash \bar{\Omega}_1$, $Q = \Omega \times (0, T)$, $\Gamma_j = \partial\Omega_j$. In general, when a lowercase letter is used for a certain quantity in the original problem, the corresponding capital letter will be used for the same quantity in the symmetrized problem.
- If f is a measurable function defined in ω, f_* denotes the decreasing rearrangement of f:

$$f_*(s) = \operatorname{Inf}\{\xi \in \mathbb{R}: |x: f(x) > \xi| \leqq s\},$$

defined for $s \in \bar{\omega}_* = [0, |\omega|]$ ($\omega_* = (0, |\omega|)$), while $\mathbf{f}$ denotes the rearrangement of f, defined in Ω, that decreases along radii:

$$\mathbf{f}(x) = f_*(\sigma_N |x|^N - m_1),$$

where $m_1 = |\omega_1|$, $x \in \Omega$. If f is defined in q, we consider its rearrangements with respect to the space variable: $f_*(s, t) = (f(\cdot, t))_*(s)$, $\mathbf{f}(x, t) = f_*(\sigma_N |x|^N - m_1, t)$.

Now, the symmetrized problem corresponding to (1.1) is

$$\widetilde{(1.1)} \qquad \begin{aligned} &\frac{\partial H}{\partial t} - \Delta\Theta = 0 \quad \text{in } Q, \\ &\Theta = G \quad \text{on } \Sigma = \partial\Omega \times (0, T), \\ &H|_{t=0} = \mathbf{h}_0, \\ &H \in a(\Theta) \quad \text{a.e. in } Q, \end{aligned}$$

where

$$G = \begin{cases} 0 & \text{on } \Sigma_0 = \Gamma_0 \times (0, T) = \partial\Omega_0 \times (0, T), \\ 1 & \text{on } \Sigma_1 = \Gamma_1 \times (0, T) = \partial\Omega_1 \times (0, T). \end{cases}$$

The problem corresponding to $(1.1)_\varepsilon$ is

$$(\widetilde{1.1})_\varepsilon \qquad \begin{aligned} &\frac{\partial H_\varepsilon}{\partial t} - \Delta \Theta_\varepsilon = 0 \quad \text{in } Q, \\ &\Theta_\varepsilon = G \quad \text{on } \Sigma, \\ &H_\varepsilon|_{t=0} = \mathbf{h}_{\varepsilon_0}, \\ &H_\varepsilon = a_\varepsilon(\Theta_\varepsilon) \quad \text{in } Q. \end{aligned}$$

Our main technical tool is the following result.

THEOREM 1. *For classical solutions of* $(1.1)_\varepsilon$ *and* $(\widetilde{1.1})_\varepsilon$, *we have the comparison*

$$(2.1)_\varepsilon \qquad \int_0^s h_{\varepsilon *}(\sigma, t)\, d\sigma - \int_0^t \int_{\gamma_1} \frac{\partial \theta_\varepsilon}{\partial \nu}\, d\gamma\, d\tau \leqq \int_0^s H_{\varepsilon *}(\sigma, t)\, d\sigma - \int_0^t \int_{\Gamma_1} \frac{\partial \Theta_\varepsilon}{\partial \nu}\, d\gamma\, d\tau$$

for every (s, t) *in* $\bar{q}_* = \bar{\omega}_* \times [0, T] = [0, |\omega|] \times [0, T]$ $(= \bar{Q}_*)$. *Here* $\partial/\partial \nu$ *denotes the outward normal derivative, and* $d\gamma$ *the* $(N-1)$*-dimensional Lebesgue measure of* $\partial\omega$ *and* $\partial\Omega$.

Similarly, for weak solutions of (1.1) *and* $(\widetilde{1.1})$ (*see* (1.3)),

$$(2.1) \qquad \int_0^s h_*(\sigma, t)\, d\sigma - \int_0^t \int_{\gamma_1} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau \leqq \int_0^s H_*(\sigma, t)\, d\sigma - \int_0^t \int_{\Gamma_1} \frac{\partial \Theta}{\partial \nu}\, d\gamma\, d\tau$$

for almost every $t \in (0, T)$ *and every* $s \in \bar{\omega}_*$.

Remarks. (1) In (2.1), the terms involving $\partial\theta/\partial\nu$ and $\partial\Theta/\partial\nu$ must be interpreted in a weak sense (see (4.10) below) because we are not guaranteed enough regularity for $\partial\theta/\partial\nu$ and $\partial\Theta/\partial\nu$ to make classical sense.

(2) Some isoperimetric inequalities involving a rearrangement and a "capacity term" (such as $\int_{\gamma_1} (\partial\theta/\partial\nu)\, d\gamma$ in (2.1)) have been previously obtained for elliptic problems in doubly connected domains, with boundary conditions 0 and 1 (respectively) on the two components of the boundary (see [7, p. 62] and [2, p. 168]).

The proof of Theorem 1 appears in §§ 3 and 4. We now give some other isoperimetric inequalities that have more physical significance and therefore can be viewed as the main results of the paper. They are all simple consequences of Theorem 1, and are proved in § 5.

If the solutions (θ, h) and (Θ, H) are "good enough," this meaning in particular that the sets $\{x \in \omega\colon h(x, \tau) > h_*(s, \tau)\}$ and $\{x \in \Omega\colon H(x, \tau) > H_*(s, \tau)\}$ have regular boundaries (in ω and Ω): $\gamma(s, \tau) = \{x \in \omega\colon h(x, \tau) = h_*(s, \tau)\}$ and $\Gamma(s, t) = \{x \in \Omega\colon H(x, \tau) = H_*(s, \tau)\}$ for almost every $\tau \in (0, t)$, then (2.1) can also be written:

$$(2.2) \qquad -\int_0^t \int_{\gamma(s,\tau)} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau \geqq -\int_0^t \int_{\Gamma(s,\tau)} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau$$

(the normal derivatives being directed outward from the sets mentioned above). The members of (2.2) are nonnegative and have the physical interpretation of being the total heat flows during the time interval $(0, t)$ from the warmest parts of volume s of ω and Ω, respectively, into the complementary colder parts. Note that these parts of ω and Ω change continuously with time.

In the particular cases $s = 0$ and $s = |\omega|$, (2.1) reduces to

$$(2.3) \qquad \int_0^t \int_{\gamma_1} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau \geqq \int_0^t \int_{\Gamma_1} \frac{\partial \Theta}{\partial \nu}\, d\gamma\, d\tau \qquad (\geqq 0),$$

$$(2.4) \qquad -\int_0^t \int_{\gamma_0} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau \geqq -\int_0^t \int_{\Gamma_0} \frac{\partial \Theta}{\partial \nu}\, d\gamma\, d\tau \qquad (\geqq 0)$$

for almost every $t \in (0, T)$. (Equation $(2.1)_\varepsilon$ reduces similarly.)

Another consequence of (2.1) is that, for any measurable set $E(t) \subset \omega$, with $|E(t)| = s$,

$$\int_0^t \int_{\gamma_1} \frac{\partial \theta}{\partial \nu} d\gamma\, d\tau - \int_{E(t)} h(x,t)\, dx \geqq \int_0^t \int_{\Gamma_1} \frac{\partial \Theta}{\partial \nu} d\gamma\, d\tau - \int_{m_1 < \sigma_N |x|^N < m_1 + s} H(x,t)\, dx. \tag{2.5}$$

In the case of a one-phase problem, let us say with $\lambda = 0$, the quantity

$$t' = \sup \left\{ t \in [0, T]: \int_0^\tau \int_{\gamma_0} \frac{\partial \theta}{\partial \nu} d\gamma\, d\tau' = 0, \text{ a.e. } \tau \in (0, t) \right\} \tag{2.6}$$

can be interpreted as the first instant at which the liquid phase reaches γ_0 (when $\lambda > 0$, $t' = 0$). With the similar definition for the symmetrized problem, T' also appears to be the time at which all the solid has melted.

We then obtain, for weak solutions of (1.1) and ($\widetilde{1.1}$), some noteworthy comparisons.

THEOREM 2. *With the above definitions*

$$t' \leqq T', \tag{2.7}$$

$$\int_s^{|\omega|} h_*(\sigma, t)\, d\sigma \geqq \int_s^{|\omega|} H_*(\sigma, t)\, d\sigma \tag{2.8}$$

for almost every $t \in (0, t')$, *and all* $s \in \bar{\omega}_*$. *More generally,*

$$\int_s^{|\omega|} \Phi(h_*(\sigma, t))\, d\sigma \geqq \int_s^{|\omega|} \Phi(H_*(\sigma, t))\, d\sigma \tag{2.9}$$

for almost every $t \in (0, t')$, *and all* $s \in \bar{\omega}_*$, *where* Φ *is any concave nondecreasing function. This implies*

$$\operatorname*{ess\,inf}_{\omega} h(t) \geqq \operatorname*{ess\,inf}_{\Omega} H(t) \quad \text{a.e. } t \in (0, t'), \tag{2.10}$$

$$\|(h(t) - \beta)_-\|_{L^p(\omega)} \leqq \|(H(t) - \beta)_-\|_{L^p(\Omega)} \quad \text{a.e. } t \in (0, t') \tag{2.11}$$

for every $\beta \in \mathbb{R}$ *and every* $p \in [1, \infty]$, *and*

$$|\{x \in \omega: h(x,t) = -\alpha\}| \leqq |\{x \in \Omega: H(x,t) = -\alpha\}| \quad \text{a.e. } t \in (0, t'). \tag{2.12}$$

The latter inequality expresses that the volume of the solid remains greater in spherical geometry (up to time t'*).*

3. Proof of (2.1)$_\varepsilon$. Let $(\theta_\varepsilon, h_\varepsilon)$ be the (unique) classical solution of $(1.1)_\varepsilon$ (cf. [6]). By the maximum principle, $0 \leqq \theta_\varepsilon \leqq 1$ in all q. Let $t \in (0, T)$ be fixed. Then, for any $\eta \in (0, 1)$, we have, by $(1.1)_\varepsilon$,

$$0 = \int_\omega \left(\frac{\partial h_\varepsilon}{\partial t} - \Delta \theta_\varepsilon \right) (\theta_\varepsilon - \eta)_+\, dx$$

$$= \int_{\theta_\varepsilon > \eta} \frac{\partial h_\varepsilon}{\partial t} (\theta_\varepsilon - \eta)\, dx + \int_{\theta_\varepsilon > \eta} |\nabla \theta_\varepsilon|^2\, dx - \int_{\gamma_1} \frac{\partial \theta_\varepsilon}{\partial \nu} (\theta_\varepsilon - \eta)\, d\gamma.$$

As (see, e.g., [7, p. 9])

$$\frac{d}{d\eta}\int_{\theta_\varepsilon>\eta}\frac{\partial h_\varepsilon}{\partial t}(\theta_\varepsilon-\eta)\,dx=-\int_{\theta_\varepsilon>\eta}\frac{\partial h_\varepsilon}{\partial t}\,dx,$$

we obtain

$$0=-\int_{\theta_\varepsilon>\eta}\frac{\partial h_\varepsilon}{\partial t}\,dx+\frac{d}{d\eta}\int_{\theta_\varepsilon>\eta}|\nabla\theta_\varepsilon|^2\,dx+\int_{\gamma_1}\frac{\partial\theta_\varepsilon}{\partial\nu}\,d\gamma. \tag{3.1}$$

Let $\mu(\eta)=|\{x\in\omega\colon \theta_\varepsilon(x,t)>\eta\}|$. Using standard rearrangement techniques (see [7]), we get, for almost every $\eta\in(0,1)$,

$$N^2\sigma_N^{2/N}(m_1+\mu(\eta))^{2-(2/N)}\leqq\left[\frac{d}{d\eta}\int_{\theta_\varepsilon>\eta}|\nabla\theta_\varepsilon|\,dx\right]^2\leqq\mu'(\eta)\frac{d}{d\eta}\int_{\theta_\varepsilon>\eta}|\nabla\theta_\varepsilon|^2\,dx. \tag{3.2}$$

Here the first inequality is the isoperimetric inequality relating the volume of the set $\omega_1\cup\{\theta_\varepsilon>\eta\}$ to its perimeter, the latter taken in the sense of De Giorgi (see, e.g., [7] for details). The second inequality is obtained from the Cauchy-Schwarz inequality applied to the difference quotients corresponding to the derivatives after passing to the limit. Combining (3.1) with (3.2) gives

$$N^2\sigma_N^{2/N}(m_1+\mu(\eta))^{2-(2/N)}\leqq-\mu'(\eta)\left[\int_{\gamma_1}\frac{\partial\theta_\varepsilon}{\partial\nu}\,d\gamma-\int_{\theta_\varepsilon>\eta}\frac{\partial h_\varepsilon}{\partial t}\,dx\right] \tag{3.3}$$

for almost every $\eta\in(0,1)$.

Next define

$$k(s,t)=\int_0^s h_{\varepsilon *}(\sigma,t)\,d\sigma. \tag{3.4}$$

Using results from the theory of relative rearrangement (see [8, Thm. 1.2, p. 60; proof of (2.12), p. 67]), we obtain for almost every η (namely, those for which $|\theta_\varepsilon=\eta|=0$),

$$\begin{aligned}\int_{\theta_\varepsilon>\eta}\frac{\partial h_\varepsilon}{\partial t}\,dx&=\int_{h_\varepsilon>a_\varepsilon(\eta)}\frac{\partial h_\varepsilon}{\partial t}\,dx\\&=\int_0^{\mu(\eta)}\left(\frac{\partial h_\varepsilon}{\partial t}\right)_{*h_\varepsilon}(\sigma,t)\,d\sigma\\&=\int_0^{\mu(\eta)}\frac{\partial h_{\varepsilon *}}{\partial t}(\sigma,t)\,d\sigma\\&=\frac{\partial k}{\partial t}(\mu(\eta),t).\end{aligned} \tag{3.5}$$

(It should be noted that the rigorous proof of (3.5) is one of the most difficult results in [8]. As it is quite long, we do not repeat it here. We just recall that $(\partial h_\varepsilon/\partial t)_{*h_\varepsilon}$ is called the relative rearrangement of $\partial h_\varepsilon/\partial t$ with respect to h_ε, and may be conceived of as the directional derivative of the map $f\to f_*$, taken at the point h_ε, in the direction $\partial h_\varepsilon/\partial t$.)

Thus (3.3) becomes

$$N^2\sigma_N^{2/N}(m_1+\mu(\eta))^{2-(2/N)}\leqq-\mu'(\eta)\left[\int_{\gamma_1}\frac{\partial\theta_\varepsilon}{\partial\nu}\,d\gamma-\frac{\partial k}{\partial t}(\mu(\eta),t)\right] \tag{3.6}$$

(for almost every $\eta \in (0, 1)$). Set (for $t \in (0, T)$ still fixed)

$$F(s) = \int_{\gamma_1} \frac{\partial \theta_\varepsilon}{\partial \nu}\, d\gamma - \frac{\partial k}{\partial t}(s, t). \tag{3.7}$$

Then F is a continuous function on $\bar{\omega}_*$ because $\partial k/\partial t(s, t) = \int_0^s (\partial h_{\varepsilon *}/\partial t)(\sigma, t)\, d\sigma = \int_0^s (\partial h_\varepsilon/\partial t)_{*h_\varepsilon}(\sigma, t)\, d\sigma$ and $(\partial h_\varepsilon/\partial t)_{*h_\varepsilon}(\sigma, t)$ is integrable as a function of σ (see [8]). We will now prove that F is also nonnegative on $\bar{\omega}_*$.

By (3.6), $(F \circ \mu)(\eta) \geqq 0$ for almost every $\eta \in (0, 1)$. Since μ is continuous from the right, so is $F \circ \mu$. It follows that $F \circ \mu \geqq 0$ on $[0, 1)$, and then on $[0, 1]$, since $(F \circ \mu)(1) = F(0) = \int_{\gamma_1} (\partial \theta_\varepsilon/\partial \nu)\, d\gamma \geqq 0$. It follows next that $F \circ \bar{\mu} \geqq 0$ on $(0, 1]$, where $\bar{\mu}(\eta) = |\theta_\varepsilon \geqq \eta|$, for $\mu(\eta - \delta) \to \bar{\mu}(\eta)$ as $\delta \downarrow 0$. Furthermore,

$$(F \circ \bar{\mu})(0) = F(|\omega|) = \int_{\gamma_1} \frac{\partial \theta_\varepsilon}{\partial \nu}\, d\gamma - \int_0^{|\omega|} \frac{\partial h_{\varepsilon *}}{\partial t}(\sigma, t)\, d\sigma.$$

This last integral also equals $\int_\omega (\partial h_\varepsilon/\partial t)\, dx = \int_\omega \Delta \theta_\varepsilon\, dx = \int_{\partial \omega} (\partial \theta_\varepsilon/\partial \nu)\, d\gamma$ (see [8]; also compare (3.5)), so

$$(F \circ \bar{\mu})(0) = -\int_{\gamma_0} \frac{\partial \theta_\varepsilon}{\partial \nu}\, d\gamma \geqq 0.$$

Thus $F \circ \bar{\mu} \geqq 0$ on all $[0, 1]$. Now let $s \in \bar{\omega}_*$, and set $\eta = \theta_{\varepsilon *}(s)$, $s' = \mu(\eta)$, $s'' = \bar{\mu}(\eta)$. Then $s' \leqq s \leqq s''$, and we will prove that $F \geqq 0$ on $[s', s'']$. We have $\partial F/\partial s = -(\partial h_{\varepsilon *}/\partial t)$. Since $\theta_\varepsilon \in H^1(0, T; L^2(\omega))$ (see [4] or § 4 below), $\theta_{\varepsilon *} \in H^1(0, T; L^2(\omega_*))$ and $\partial h_{\varepsilon *}/\partial t = a_\varepsilon'(\theta_{\varepsilon *})\, \partial \theta_{\varepsilon *}/\partial t$ is almost everywhere constant on (s', s'') (see [8, pp. 60, 63]). Therefore F is affine on $[s', s'']$, and since $F(s')$ and $F(s'')$ were shown to be nonnegative, we conclude that $F \geqq 0$ on $[s', s'']$. In particular $F(s) \geqq 0$ and since $s \in \bar{\omega}_*$ was arbitrary this proves that $F \geqq 0$ on $\bar{\omega}_*$.

Now (3.3) can be written

$$1 \leqq -N^{-2} \sigma_N^{-(2/N)} (m_1 + \mu(\eta))^{(2/N)-2} F(\mu(\eta)) \mu'(\eta) \quad \text{a.e. } \eta \in (0, 1).$$

Using that $\mu(\eta)$ is a nonincreasing function, integration from η to η', where $0 \leqq \eta < \eta' \leqq 1$, gives

$$\eta' - \eta \leqq N^{-2} \sigma_N^{-(2/N)} \int_{\mu(\eta')}^{\mu(\eta)} (m_1 + s)^{(2/N)-2} F(s)\, ds.$$

As in [7, pp. 24, 31] and [8], this shows that

$$N^{-2} \sigma_N^{-(2/N)} (m_1 + s)^{(2/N)-2} F(s) + \frac{\partial \theta_{\varepsilon *}}{\partial s} \geqq 0 \quad \text{a.e. } s \in \omega_*. \tag{3.8}$$

Now we take the time-dependence into account. Set

$$y(s, t) = \int_0^s h_{\varepsilon *}(\sigma, t)\, d\sigma - \int_0^t \int_{\gamma_1} \frac{\partial \theta_\varepsilon}{\partial \nu}(x, \tau)\, d\gamma\, d\tau \tag{3.9}$$

for $(s, t) \in \bar{\omega}_* \times [0, T]$. Then

$$\frac{\partial y}{\partial t}(s, t) = -F(s), \qquad \frac{\partial y}{\partial s}(s, t) = h_{\varepsilon *}(s, t).$$

Since $\theta_\varepsilon = b_\varepsilon(h_\varepsilon)$, where $b_\varepsilon = a_\varepsilon^{-1}$ is strictly increasing, we also have

$$\theta_{\varepsilon *} = b_\varepsilon(h_{\varepsilon *}) = b_\varepsilon\left(\frac{\partial y}{\partial s}\right).$$

Therefore (3.8) shows that y satisfies

$$N^{-2} \sigma_N^{-(2/N)} (m_1 + s)^{(2/N)-2} \frac{\partial y}{\partial t} - \frac{\partial}{\partial s}\left(b_\varepsilon\left(\frac{\partial y}{\partial s}\right)\right) \leqq 0 \quad \text{a.e. in } q_* = \omega_* \times (0, T), \tag{3.10}$$

i.e., y is a subsolution of a parabolic equation. Moreover, we get the following boundary and initial conditions for y:

$$\frac{\partial y}{\partial s}(0,t)=a_\varepsilon(1), \qquad \frac{\partial y}{\partial s}(|\omega|,t)=a_\varepsilon(0), \tag{3.11}$$
$$y(s,0)=y_{\varepsilon_0}(s),$$

where $y_{\varepsilon_0}(s)=\int_0^s h_{\varepsilon_0 *}(\sigma)\,d\sigma$.

For the symmetrized problem $(\widetilde{1.1})_\varepsilon$ we obtain as in [8], for

$$Y(s,t)=\int_0^s H_{\varepsilon *}(\sigma,t)\,d\sigma-\int_0^t\int_{\Gamma_1}\frac{\partial\Theta_\varepsilon}{\partial\nu}(x,\tau)\,d\gamma\,d\tau, \tag{$\widetilde{3.9}$}$$

$$N^{-2}\sigma_N^{-(2/N)}(m_1+s)^{(2/N)-2}\frac{\partial Y}{\partial t}-\frac{\partial}{\partial s}\left(b_\varepsilon\left(\frac{\partial Y}{\partial s}\right)\right)=0 \quad \text{in } q_*(=Q_*), \tag{$\widetilde{3.10}$}$$

$$\frac{\partial Y}{\partial s}(0,t)=a_\varepsilon(1), \qquad \frac{\partial Y}{\partial s}(|\omega|,t)=a_\varepsilon(0), \tag{$\widetilde{3.11}$}$$
$$Y(s,0)=y_{\varepsilon_0}(s)$$

(observe that $(\mathbf{h}_{\varepsilon_0})_*=h_{\varepsilon_0 *}$). In fact, since Θ_ε, H_ε, are radially symmetric and decreasing, the first line in $(\widetilde{1.1})_\varepsilon$ can be written:

$$\frac{\partial H_{\varepsilon *}}{\partial t}-\frac{\partial}{\partial s}\left(N^2\sigma_N^{(2/N)}(m_1+s)^{2-(2/N)}\frac{\partial\Theta_{\varepsilon *}}{\partial s}\right)=0 \quad \text{in } q_*.$$

Now, we get $(\widetilde{3.10})$ by integrating between 0 and s, noting that

$$\frac{\partial H_{\varepsilon *}}{\partial t}=\frac{\partial^2 Y}{\partial s\,\partial t} \quad \text{and} \quad \int_{\Gamma_1}\frac{\partial\Theta_\varepsilon}{\partial\nu}\,d\gamma=-N^2\sigma_N^{(2/N)}m_1^{2-(2/N)}\frac{\partial\Theta_{\varepsilon *}}{\partial s}(0).$$

Now $(2.1)_\varepsilon$ of Theorem 1 simply states that $y\leqq Y$ in $\bar{q}_*$. To prove this inequality, we multiply the difference between (3.10) and $(\widetilde{3.10})$ by $(y-Y)_+$ and integrate with respect to s for fixed t. Taking the boundary conditions for y and Y into account and using that b_ε is monotone increasing, we get

$$\begin{aligned}
0&\geqq\int_0^{|\omega|}(m_1+s)^{(2/N)-2}\frac{\partial(y-Y)}{\partial t}(y-Y)_+\,ds\\
&\quad-N^2\sigma_N^{2/N}\int_0^{|\omega|}\frac{\partial}{\partial s}\left[b_\varepsilon\left(\frac{\partial y}{\partial s}\right)-b_\varepsilon\left(\frac{\partial Y}{\partial s}\right)\right](y-Y)_+\,ds\\
&=\frac{1}{2}\int_0^{|\omega|}(m_1+s)^{(2/N)-2}\frac{\partial}{\partial t}(y-Y)_+^2\,ds\\
&\quad+N^2\sigma_N^{2/N}\int_0^{|\omega|}\left[b_\varepsilon\left(\frac{\partial y}{\partial s}\right)-b_\varepsilon\left(\frac{\partial Y}{\partial s}\right)\right]\frac{\partial}{\partial s}(y-Y)_+\,ds\\
&=\frac{1}{2}\frac{d}{dt}\int_0^{|\omega|}(m_1+s)^{(2/N)-2}(y-Y)_+^2\,ds\\
&\quad+N^2\sigma_N^{2/N}\int_{y>Y}\left[b_\varepsilon\left(\frac{\partial y}{\partial s}\right)-b_\varepsilon\left(\frac{\partial Y}{\partial s}\right)\right]\left[\frac{\partial y}{\partial s}-\frac{\partial Y}{\partial s}\right]ds\\
&\geqq\frac{1}{2}\frac{d}{dt}\int_0^{|\omega|}(m_1+s)^{(2/N)-2}(y-Y)_+^2\,ds.
\end{aligned}$$

Since $y = Y$ for $t = 0$, it follows that

$$\int_0^{|\omega|} (m_1 + s)^{(2/N)-2} (y - Y)_+^2 \, ds \leqq 0 \quad \text{for } t \geqq 0,$$

and hence that $y \leqq Y$ in $\bar{q}_*$ as desired.

4. Proof of (2.1). The weak solution (θ, h) of (1.1) is obtained as the limit of the solution $(\theta_\varepsilon, h_\varepsilon)$ of $(1.1)_\varepsilon$ as $\varepsilon \to 0$, and we will accordingly obtain (2.1) by letting $\varepsilon \to 0$ in $(2.1)_\varepsilon$. For convenience we review part of the construction of (θ, h).

Let $(\theta_\varepsilon, h_\varepsilon)$ be the solution of $(1.1)_\varepsilon$, and let $q_t = \omega \times (0, t)$ for $t \in (0, T)$. Then we have

$$\begin{aligned} 0 &= \int\int_{q_t} \frac{\partial \theta_\varepsilon}{\partial \tau} \left(\frac{\partial h_\varepsilon}{\partial \tau} - \Delta \theta_\varepsilon \right) dx \, d\tau \\ &= \int\int_{q_t} \frac{\partial \theta_\varepsilon}{\partial \tau} \frac{\partial h_\varepsilon}{\partial \tau} \, dx \, dt + \int_0^t \left[\int_\omega \nabla \frac{\partial \theta_\varepsilon}{\partial \tau} \nabla \theta_\varepsilon \, dx - \int_{\partial \omega} \frac{\partial \theta_\varepsilon}{\partial \tau} \frac{\partial \theta_\varepsilon}{\partial \nu} \, d\gamma \right] d\tau \end{aligned}$$

or

$$\int\int_{q_t} a_\varepsilon'(\theta_\varepsilon) \left| \frac{\partial \theta_\varepsilon}{\partial \tau} \right|^2 dx \, d\tau + \frac{1}{2} \int_\omega |\nabla \theta_\varepsilon|^2 \, dx = \frac{1}{2} \int_\omega |\nabla \theta_{\varepsilon_0}|^2 \, dx.$$

By assumptions (1.4) and (1.7), when $\varepsilon \to 0$, $\int_\omega |\nabla \theta_{\varepsilon_0}|^2 \, dx$ is bounded, and a_ε' is bounded from below. Hence

$$\int\int_{q_t} \left| \frac{\partial \theta_\varepsilon}{\partial \tau} \right|^2 dx \, d\tau \leqq C, \tag{4.1}$$

$$\int_\omega |\nabla \theta_\varepsilon(x, t)|^2 \, dx \leqq C, \tag{4.2}$$

for some constant C independent of ε and t. By the maximum principle, the families $\{\theta_\varepsilon\}_{\varepsilon > 0}$ and $\{h_\varepsilon\}_{\varepsilon > 0}$ are bounded in $L^\infty(q)$, and, by the above estimates, it then follows that $\{\theta_\varepsilon\}_{\varepsilon > 0}$ is bounded in $L^\infty(0, T; H^1(\omega))$ and in $H^1(q)$. By repeated extraction of subsequences from $\{\varepsilon\}$ we thus can find a sequence ε_n such that

$$\begin{aligned} &\theta_{\varepsilon_n} \to \theta \text{ weakly* in } L^\infty(q) \text{ and } L^\infty(0, T; H^1(\omega)), \\ &\qquad \text{weakly in } H^1(q), \\ &\qquad \text{strongly in } L^2(q) \text{ (by compactness)}, \\ &h_{\varepsilon_n} \to h \text{ weakly* in } L^\infty(q) \end{aligned} \tag{4.3}$$

for some pair (θ, h) satisfying

$$\theta \in L^\infty(q) \cap H^1(q) \cap L^\infty(0, T; H^1(\omega)), \qquad h \in L^\infty(q). \tag{4.4}$$

In the following, we shall replace ε_n by ε for simplicity.

Now (θ, h) is the required weak solution. In fact, it follows immediately that (1.3) holds, since $(\theta_\varepsilon, h_\varepsilon)$ satisfies the same equation with h_0 replaced by h_{ε_0}. To check that $h \in a(\theta)$ almost everywhere, it is enough (since a is maximally monotone) to check that

$$\langle h' - h, \theta' - \theta \rangle \geqq 0 \tag{4.5}$$

(where $\langle f, g\rangle = \iint_q fg\,dx\,dt$) for all θ', $h' \in L^\infty(q)$ satisfying $h' \in a(\theta')$ almost everywhere. Writing

$$\langle h'-h, \theta'-\theta\rangle = \langle h_\varepsilon - h, \theta'-\theta\rangle + \langle h'-h_\varepsilon, \theta_\varepsilon-\theta\rangle + \langle h'-h_\varepsilon, \theta'-\theta_\varepsilon\rangle, \tag{4.6}$$

we have (as $\varepsilon \to 0$)

$$\langle h_\varepsilon - h, \theta'-\theta\rangle \to 0 \tag{4.7}$$

since $h_\varepsilon \to h$ weakly* in $L^\infty(q)$ (by (4.3)), and

$$|\langle h'-h_\varepsilon, \theta_\varepsilon-\theta\rangle| \leqq \|h'-h_\varepsilon\|_{L^2(q)} \|\theta_\varepsilon-\theta\|_{L^2(q)} \to 0, \tag{4.8}$$

since $\theta_\varepsilon \to \theta$ in $L^2(q)$, and $\|h'-h_\varepsilon\|_{L^2(q)}$ is bounded (by (4.3)).

Remark. This is the only place where we use the fact that $b = a^{-1}$ is Lipschitz continuous. This property of a makes it possible to choose a_ε satisfying (1.4), from which we get the estimate (4.1) for $\partial\theta_\varepsilon/\partial t$; then weak convergence in $H^1(q)$ and strong convergence in $L^2(q)$ of θ_ε to θ follows. Observe that (4.2) holds independently of (1.4).

For the last term in (4.6), we have

$$\begin{aligned}\langle h'-h_\varepsilon, \theta'-\theta_\varepsilon\rangle &= \langle h'-h_\varepsilon, b(h')-b_\varepsilon(h_\varepsilon)\rangle \\ &= \langle h'-h_\varepsilon, b_\varepsilon(h')-b_\varepsilon(h_\varepsilon)\rangle + \langle h'-h_\varepsilon, b(h')-b_\varepsilon(h')\rangle.\end{aligned}$$

The first bracket in the last member is nonnegative as b_ε is nondecreasing, and the second one tends to zero with ε, as b_ε converges uniformly to b:

$$|\langle h'-h_\varepsilon, b(h')-b_\varepsilon(h')\rangle| \leqq \|h'-h_\varepsilon\|_{L^\infty(q)} \int_q |b(h')-b_\varepsilon(h')|\,dx\,dt.$$

Then

$$\varliminf_{\varepsilon\to 0} \langle h'-h_\varepsilon, \theta'-\theta_\varepsilon\rangle \geqq 0. \tag{4.9}$$

Now (4.5) follows by combining (4.7)-(4.9). Thus (θ, h) is a weak solution.

We now pass to the limit in $(2.1)_\varepsilon$. First, we have to give a weak interpretation of the two members of (2.1), since the regularity of θ that we have is not enough for $\partial\theta/\partial\nu$ and $\partial\Theta/\partial\nu$ to make classical sense.

Let φ be an arbitrary smooth function in ω (say, $\varphi \in \mathscr{C}^2(\bar\omega)$) with boundary values $\varphi = 0$ on γ_0, $\varphi = 1$ on γ_1. Then for (θ, h), a "good enough" solution of $(\partial h/\partial t) - \Delta\theta = 0$,

$$\begin{aligned}\int_0^t \int_{\gamma_1} \frac{\partial\theta}{\partial\nu}\,d\gamma\,d\tau &= \int_0^t \int_{\partial\omega} \varphi \frac{\partial\theta}{\partial\nu}\,d\gamma\,d\tau \\ &= \int_0^t \int_\omega \nabla\varphi\,\nabla\theta\,dx\,d\tau + \int_0^t \int_\omega \varphi\,\Delta\theta\,dx\,d\tau \\ &= \int_0^t \int_\omega \nabla\varphi\,\nabla\theta\,dx\,d\tau + \int_\omega (h(x,t)-h_0(x))\varphi(x)\,dx.\end{aligned} \tag{4.10}$$

Here the last member makes sense for almost every t for any $\theta \in L^\infty(0, T; H^1(\omega))$, $h \in L^\infty(q)$, and defines an (almost everywhere) bounded measurable function of t. Therefore, when (θ, h) is the weak solution of (1.1) satisfying (4.4), we choose the last member of (4.10) to be our definition of $\int_0^t \int_{\gamma_1} (\partial\theta/\partial\nu)\,d\gamma\,d\tau$ for almost every $t \in (0, T)$; we choose similarly for $\int_0^t \int_{\Gamma_1} \partial\Theta/\partial\nu\,d\gamma\,d\tau$. We still have to check that this definition is independent of the choice of φ. This amounts to showing that if $\varphi \in \mathscr{C}^2(\bar\omega)$ has boundary values zero on $\partial\omega$, then

$$\int_0^t \int_\omega \nabla\varphi\,\nabla\theta\,dx\,d\tau + \int_\omega (h(x,t)-h_0(x))\varphi(x)\,dx = 0 \quad \text{a.e. } t \in (0, T). \tag{4.11}$$

Now, (4.11) is an equation of the form $\int_0^t f(\tau)\,d\tau + F(t) = 0$ for almost every $t \in (0, T)$, for two functions f and F in $L^1(0, T)$. The latter equation is equivalent to $\int_0^T f(t)\psi(t)\,dt = \int_0^T F(t)\psi'(t)\,dt$ for every $\psi \in \mathscr{C}^1[0, T]$ with $\psi(T) = 0$. Thus (4.11) is equivalent to

$$\int_0^T \int_\omega \nabla\theta(x, t)\nabla\varphi(x)\psi(t)\,dx\,dt = \int_0^T \int_\omega (h(x, t) - h_0(x))\varphi(x)\psi'(t)\,dx\,dt \tag{4.12}$$

for ψ as above. Now, the truth of (4.12) follows by integration in the definition (1.3) for a weak solution, taking $\varphi(x)\psi(t)$ as a test function (this integration by parts is justified by $\theta \in L^\infty(0, T; H^1(\omega))$ for (θ, h) a weak solution of (1.1)).

For later use we remark that $\int_0^t \int_{\gamma_0} (\partial\theta/\partial\nu)\,d\gamma\,d\tau$ may be defined by the same formula (4.10), taking a test function with the exchanged boundary values, e.g., $1 - \varphi$, with φ as in (4.10). Then, for such φ's,

$$\int_0^t \int_{\gamma_0} \frac{\partial\theta}{\partial\nu}\,d\gamma\,d\tau = -\int_0^t \int_\omega \nabla\varphi\,\nabla\theta\,dx\,d\tau + \int_\omega (h(x, t) - h_0(x))(1 - \varphi(x))\,dx, \tag{4.10}'$$

and from (4.10), (4.10)',

$$\begin{aligned} \int_0^t \int_{\gamma_0} \frac{\partial\theta}{\partial\nu}\,d\gamma\,d\tau + \int_0^t \int_{\gamma_1} \frac{\partial\theta}{\partial\nu}\,d\gamma\,d\tau &= \int_\omega (h(x, t) - h_0(x))\,dx \\ &= \int_0^{|\omega|} (h_*(\sigma, t) - h_{0*}(\sigma))\,d\sigma. \end{aligned} \tag{4.13}$$

Now, with $\varphi \in \mathscr{C}^2(\bar{\omega})$ (respectively, $\Phi \in \mathscr{C}^2(\bar{\Omega})$) in (4.10), (2.1) (to be proven) becomes

$$\begin{aligned} &\int_0^t \int_\omega \nabla\theta\,\nabla\varphi\,dx\,d\tau + \int_\omega (h(x, t) - h_0(x))\varphi(x)\,dx - \int_0^s h_*(\sigma, t)\,d\sigma \\ &\qquad \geqq \int_0^t \int_\Omega \nabla\Theta\,\nabla\Phi\,dx\,d\tau + \int_\Omega (H(x, t) - \mathbf{h}_0(x))\Phi(x)\,dx - \int_0^s H_*(\sigma, t)\,d\sigma \end{aligned} \tag{4.14}$$

for $s \in \bar{\omega}_*$ and almost every $t \in (0, T)$. For fixed $s \in \bar{\omega}_*$, both members above are integrable functions of t. Therefore (4.14) is equivalent to a statement that, for every $s \in \bar{\omega}_*$, and every nonnegative function $\psi \in \mathscr{C}[0, T]$,

$$\begin{aligned} &\int_0^T \int_0^t \int_\omega \nabla\theta(x, \tau)\nabla\varphi(x)\psi(t)\,dx\,d\tau\,dt + \iint_q (h(x, t) - h_0(x))\varphi(x)\psi(t)\,dx\,dt \\ &\qquad - \int_0^T \int_0^s h_*(\sigma, t)\psi(t)\,d\sigma\,dt \\ &\qquad \geqq \int_0^T \int_0^t \int_\Omega \nabla\Theta(x, \tau)\nabla\Phi(x)\psi(t)\,dx\,d\tau\,dt \\ &\qquad\qquad + \iint_Q (H(x, t) - \mathbf{h}_0(x))\Phi(x)\psi(t)\,dx\,dt \\ &\qquad\qquad - \int_0^T \int_0^s H_*(\sigma, t)\psi(t)\,d\sigma\,dt \end{aligned} \tag{4.15}$$

(we have to make this extra integration because the convergence h_ε to h is not established for fixed t (see (4.3)), as we would need to get (4.14) from the corresponding $(4.14)_\varepsilon$).

By $(2.1)_\varepsilon$, we have $(4.15)_\varepsilon$, that is, (4.15) holds with θ, Θ, h, H, h_0, $\mathbf{h}_0$ replaced by θ_ε, Θ_ε, h_ε, H_ε, h_{ε_0}, $\mathbf{h}_{\varepsilon_0}$, respectively. For the first terms on the left-hand sides of

(4.15) and $(4.15)_\varepsilon$, we have

$$\int_0^T \int_0^t \int_\omega \nabla\theta_\varepsilon(x,\tau)\nabla\varphi(x)\psi(t)\,dx\,d\tau\,dt \to \int_0^T \int_0^t \int_\omega \nabla\theta(x,\tau)\nabla\varphi(x)\psi(t)\,dx\,d\tau\,dt$$

as $\varepsilon \to 0$, because $\theta_\varepsilon \to \theta$ weakly* in $L^\infty(0, T; H^1(\omega))$ (see (4.3)). We have similar results for the first terms on the right-hand sides of (4.15) and $(4.15)_\varepsilon$. As for the second terms on the left-hand sides,

$$\iint_q (h_\varepsilon(x,t) - h_{\varepsilon_0}(x))\varphi(x)\psi(t)\,dx\,dt \to \iint_q (h(x,t) - h_0(x))\varphi(x)\psi(t)\,dx\,dt$$

because $h_\varepsilon \to h$ weakly* in $L^\infty(q)$ (see (4.3)), and $h_{\varepsilon_0} \to h_0$ in $L^1(q)$ (see (1.6)). Similarly, for the right-hand sides, ($\mathbf{h}_{\varepsilon_0} \to \mathbf{h}_0$ in $L^1(Q)$ follows from $h_{\varepsilon_0} \to h_0$ because the rearrangement operator is a contraction in L^p-spaces (see [7, pp. 7, 8])). For the last terms on the left-hand sides we have

$$\underline{\lim}_{\varepsilon\to 0} \int_0^T \int_0^s h_{\varepsilon_*}(\sigma,t)\psi(t)\,d\sigma\,dt \geqq \int_0^T \int_0^s h_*(\sigma,t)\psi(t)\,d\sigma\,dt \tag{4.16}$$

(the map $h \to \int_0^T \int_0^s h_*(\sigma,t)\psi(t)\,d\sigma\,dt$ is weakly lower semicontinuous $L^1(q) \to \mathbb{R}$). In fact, for every measurable subset $E(t) \subset \omega$, with $|E(t)| = s$, by the Hardy-Littlewood inequality

$$\int_0^s h_{\varepsilon_*}(\sigma,t)\,d\sigma \geqq \int_{E(t)} h_\varepsilon(x,t)\,dx;$$

hence

$$\int_0^T \int_0^s h_{\varepsilon_*}(\sigma,t)\psi(t)\,d\sigma\,dt \geqq \int_0^T \int_{E(t)} h_\varepsilon(x,t)\psi(t)\,dx\,dt,$$

and, if we let $\varepsilon \to 0$,

$$\underline{\lim}_{\varepsilon\to 0} \int_0^T \int_0^s h_{\varepsilon_*}(\sigma,t)\psi(t)\,d\sigma\,dt \geqq \int_0^T \int_{E(t)} h(x,t)\psi(t)\,dx\,dt$$

(since $h_\varepsilon \to h$ weakly in $L^1(q)$ by (4.3)). Now, $E(t)$ can be chosen such that $|E(t) = s|$ and

$$\int_{E(t)} h(x,t)\,dx = \int_0^s h_*(\sigma,t)\,d\sigma,$$

which proves (4.16). Finally we consider the last terms in the right-hand sides of (4.15) and $(4.15)_\varepsilon$. It is clear (by uniqueness of solutions) that Θ_ε, Θ, H_ε, and H are radially symmetric, i.e., are functions of $r = |x|$ and t only. Moreover, Θ_ε and H_ε are nonincreasing as functions of $|x|$ for fixed t (i.e., $\Theta_\varepsilon(x,t) \geqq \Theta_\varepsilon(y,t)$ whenever $|x| \leqq |y|$), as can be seen by applying the parabolic maximum principle to $\partial\Theta_\varepsilon/\partial r$ and $\partial H_\varepsilon/\partial r$ (this pair of functions satisfies a parabolic equation). It also follows that Θ and H are nonincreasing as functions of $|x|$ because the set of such functions (nonincreasing in $|x|$) is closed and convex in, e.g., $L^2(Q)$ and therefore weakly closed in $L^2(Q)$ (use (4.3)). From the above, we get

$$\int_0^s H_{\varepsilon_*}(\sigma,t)\,d\sigma = \int_{m_1 < \sigma_N |x|^N < m_1+s} H_\varepsilon(x,t)\,dx, \tag{4.17}$$

$$\int_0^s H_*(\sigma,t)\,d\sigma = \int_{m_1 < \sigma_N |x|^N < m_1+s} H(x,t)\,dx, \tag{4.18}$$

and since $H_\varepsilon \to H$ weakly* in $L^\infty(Q)$ (see (4.3)) we finally obtain, as $\varepsilon \to 0$,

$$\int_0^T \int_0^s H_{\varepsilon_*}(\sigma, t)\psi(t)\, d\sigma\, dt \to \int_0^T \int_0^s H_*(\sigma, t)\psi(t)\, d\sigma\, dt.$$

Now, all the above shows that (4.15) results from letting $\varepsilon \to 0$ in $(4.15)_\varepsilon$. This completes the proof of Theorem 1.

5. Proofs of (2.2)–(2.5) and (2.7)–(2.12). We first observe that, in view of (4.13), (2.1) also can be written:

$$(5.1)\quad -\int_0^t \int_{\gamma_0} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau + \int_s^{|\omega|} h_*(\sigma, t)\, d\sigma \geqq -\int_0^t \int_{\Gamma_0} \frac{\partial \Theta}{\partial \nu}\, d\gamma\, d\tau + \int_s^{|\omega|} H_*(\sigma, t)\, d\sigma.$$

By taking $s = 0$ in (2.1), and $s = |\omega|$ in (5.1), we obtain (2.3) and (2.4).

More generally, for "classical" solutions (θ, h) and (Θ, H), if $s > 0$ is such that $\gamma(s, \tau) = \{x \in \omega\colon h(x, \tau) = h_*(s, \tau)\}$ and $\Gamma(s, \tau) = \{x \in \Omega\colon H(x, \tau) = H_*(s, \tau)\}$ are regular curves (in particular, if they have measure zero) for almost every $\tau \in (0, t)$, using the technique of relative rearrangement [8] gives us

$$\begin{aligned}
\int_0^s (h_*(\sigma, t) - h_{0*}(\sigma))\, d\sigma &= \int_0^s \int_0^t \frac{\partial h_*}{\partial \tau}(\sigma, \tau)\, d\tau\, d\sigma \\
&= \int_0^t \int_0^s \frac{\partial h_*}{\partial \tau}(\sigma, \tau)\, d\sigma\, d\tau \\
&= \int_0^t \int_{h(x,\tau) > h_*(s,\tau)} \frac{\partial h}{\partial \tau}\, dx\, d\tau \\
&= \int_0^t \int_{h(x,\tau) > h_*(s,\tau)} \Delta\theta\, dx\, d\tau \\
&= \int_0^t \int_{\gamma_1} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau + \int_0^t \int_{\gamma(s,\tau)} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau
\end{aligned}$$

($\partial/\partial\nu$ the outward normal of $\{h(x, \tau) > h_*(s, \tau)\}$); we proceed similarly for (Θ, H). Thus in (2.1),

$$\int_0^t \int_{\gamma_1} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau - \int_0^s h_*(\sigma, t)\, d\tau = -\int_0^t \int_{\gamma(s,\tau)} \frac{\partial \theta}{\partial \nu}\, d\gamma\, d\tau - \int_0^s h_{0*}(\sigma)\, d\sigma,$$

and similarly for (Θ, H), which explains (2.2). (Note also that $(\mathbf{h}_0)_* = h_{0*}$.)

Now, (2.5) holds because, by the Hardy-Littlewood inequality, $\int_{E(t)} h\, dx \leqq \int_0^s h_*\, d\sigma$, when $|E(t)| = s$, and $\int_{m_1 < \sigma_N |x|^N < m_1 + s} H\, dx = \int_0^s H_*\, d\sigma$ (see (4.18)).

Next, suppose that $\int_0^t \int_{\gamma_0} \partial\theta/\partial\nu\, d\gamma\, d\tau = 0$ (the integral being at least defined by (4.10)′), for almost every $t \in (0, t')$ for some $t' > 0$. This occurs in the one-phase problem (with $\lambda = 0$) if the liquid phase does not reach γ_0 initially (because θ is identically zero in the solid phase when $\lambda = 0$). By (2.4), then also

$$\int_0^t \int_{\Gamma_0} \frac{\partial \Theta}{\partial \nu}\, d\gamma\, d\tau = 0 \quad \text{a.e. } t \in (0, t'),$$

that is, we get (2.7), and (5.1) simply reduces to (2.8) for $t \in (0, t')$.

Now (2.9) follows easily as in the convexity result [1, p. 174], and (2.10) follows from (2.8) by dividing by $|\omega| - s$ and letting s tend to $|\omega|$. Choosing $\Phi(h) = -[(h - \beta)_-]^p$ ($p \geqq 1$) in (2.9) with $s = 0$, we obtain (2.11) for $p < \infty$; letting p tend to ∞ then gives

(2.11) for $p=\infty$. Choosing $\Phi(h)=-(h-\beta)_-/(\alpha+\beta)$, $\beta>-\alpha$, in (2.9) with $s=0$, we finally obtain (2.12) by letting $\beta\to-\alpha$ (observe that $h, H\geqq-\alpha$ almost everywhere when $\lambda=0$).

Acknowledgment. The authors are grateful to J. I. Diaz for several valuable discussions on the subject of the paper.

REFERENCES

[1] C. BANDLE, *Isoperimetric Inequalities and Applications*, Pitman, Boston, London, 1980.

[2] J. I. DIAZ, *Applications of symmetric rearrangement to certain nonlinear elliptic equations with a free boundary*, in Nonlinear Differential Equations, J. K. Hale and P. Martinez-Amores, eds., Research Notes in Math., Pitman, London, 1985, pp. 155-181.

[3] A. FRIEDMAN, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 139 (1968), pp. 51-87.

[4] ———, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[5] B. GUSTAFSSON AND J. MOSSINO, *Quelques inégalités isopérimétriques pour le problème de Stefan*, C.R. Acad. Sci. Paris. Sér. I Math., 305 (1987), pp. 669-672.

[6] O. A. LADYZENSKAYA, V. SOLONNIKOV, AND N. URALCEVA, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Transl. 23, American Mathematical Society, Providence, RI, 1968.

[7] J. MOSSINO, *Inégalités Isopérimétriques et Applications en Physique*, Hermann, Paris, 1984.

[8] J. MOSSINO AND J. M. RAKOTOSON, *Isoperimetric inequalities in parabolic equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 13 (1986), pp. 51-73.

[9] O. A. OLEINIK, *On the equations of the type of nonstationary filtration*, Dokl. Akad. Nauk SSSR, 113 (1957), pp. 1210-1213.

[10a] J. L. VAZQUEZ, *Symétrisation pour $u_t=\Delta\varphi(u)$ et applications*, C.R. Acad. Sci. Paris Sér. I. Math., 295 (1982), pp. 71-74.

[10b] ———, *Erratum*, C.R. Acad. Sci. Paris Sér. I. Math., 296 (1983), pp. 455-458.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1109–1127, September 1989

006

SHORT-TIME ASYMPTOTICS OF THE HEAT KERNEL ON A CONCAVE BOUNDARY*

PEI HSU†

Abstract. A probabilistic method is used to study short-time asymptotic behavior of heat kernel in the exterior of an insulated smooth convex body. The expansion of the heat kernel $p(t, a, b)$ when both a and b are on the boundary is obtained by reducing the problem to the computation of a Wiener functional on a Brownian bridge. The leading terms of $\log p(t, a, b)$ are proved to be

$$-\frac{\rho^2}{2t} - \frac{\mu_1 \rho^{1/3}}{t^{1/3}} \int_0^\rho N(s)^{2/3} ds - \left(\frac{d}{2} + \frac{1}{6}\right) \log t + C_0 + o(1)$$

where ρ is the distance between a and b, $N(s)$ is the normal curvature of the geodesic joining a and b, and C_0 is an explicitly identified constant.

Key words. heat kernel, Laplace–Beltrami operator, normal curvature, diffusion process on manifold, Brownian bridge, Feynman–Kac formula, Girsanov formula

AMS(MOS) subject classifications. primary 58G32; secondary 35K05

1. Introduction.

Let M be the exterior of a smooth, strictly convex body in Euclidean space. Let a, b be two points on the boundary such that there is a unique distance-minimizing curve γ joining them which lies completely in M. Since ∂M is concave viewed from M, it is clear then γ must be the unique geodesic joining a and b in ∂M when ∂M is viewed as a Riemannian manifold with induced metric. Let us denote the length of γ by $\rho = d(a, b)$.

Let $p(t, x, y)$ be the heat kernel of the Laplace operator $\Delta/2$ on the domain M with the Neumann boundary condition on ∂M. In this paper we are interested in the asymptotic behavior of $p(t, a, b)$ as $t \to 0$. Recall the basic result of Varadhan [10]:

$$\lim_{t \to 0} t \log p(t, a, b) = -\frac{1}{2} \rho^2. \tag{1.1}$$

Our problem is to seek an improvement of (1.1) which reflects the geometry of the boundary near the geodesic γ. It has long been recognized in the diffraction theory that the correction to (1.1) takes the following form

$$\log p(t, a, b) = -\frac{\rho^2}{2t} - \frac{C}{t^{1/3}} + o\left(\frac{1}{t^{1/3}}\right) \tag{1.2}$$

where C is a positive constant. In fact, using the idea of path integration, Buslaev [2] was able to give a heuristic argument of (1.2) and identified constant C explicitly. However, to make his argument into a mathematically acceptable proof seems not to be a simple matter. Equation (1.2) has long been known in physics literature as Busleav's conjecture.

---

*Received by the editors November 9, 1987; accepted for publication (in revised form) October 24, 1988. The research of this author was partly supported by National Science Foundation grant DMS-86-00233.

†Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. Present address, Department of Mathematics, Northwestern University, Evanston, Illinois 60208.

We will study the expansion (1.2) by a probabilistic method initiated by Molchanov [9]. Our result can be briefly described as follows. We parametrize the geodesic γ by arclength. Let $N(s)$ be the normal curvature of γ (as a curve in $\partial M \hookrightarrow M$) at point $\gamma(s)$. $N(s)$ is simply the curvature of γ when viewed as a curve in the Euclidean space. Since ∂M is the exterior of a strictly convex body, $N(s)$ is strictly positive along γ. The asymptotic behavior of $p(t,a,b)$ is described by

$$\log p(t,a,b) = -\frac{\rho^2}{2t} - \frac{\mu_1 \rho^{1/3}}{t^{1/3}} \int_0^\rho N(s)^{2/3} ds - \left(\frac{d}{2} + \frac{1}{6}\right) \log t + C_0 + o(1).$$

Here μ_1 is the first eigenvalue of $\phi''(x)/2 - |x|\phi(x) + \mu\phi(x) = 0$ on R^1 and C_0 is nonzero constant.

Probabilistically, the heat kernel $p(t,x,y)$ is the transition density function of reflecting Brownian motion on M. By a series of asymptotic analyses, we reduce the computation of $p(t,a,b)$ to that of the following Wiener functional on the standard Brownian bridge $\hat{W}$:

$$E\left[\exp\left\{-\lambda \int_0^1 l(s)|\hat{W}_s| ds\right\}\right], \tag{1.3}$$

where l is a smooth, strictly positive function.

Our research is inspired by the work of Ikeda [5], where a special case of the present problem is discussed. In [5], manifold M is assumed to have the form of a warped product (thus the normal curvature $N(s)$ is a constant). This assumption allows us to construct Brownian motion on M by skew product and to simplify the analysis involved. In our present work, we have further explored some ideas from [5]. For a related problem under a different context, see Melrose and Taylor [8].

The plan of this work is as follows. In §2, we make precise our geometric assumptions and state our main theorem. The proof of the main theorem is outlined in §3. In order not to interrupt the main line of argument, verifications of some intermediate results used in §3 are relegated to §§4 and 5. The asymptotic analysis of the Wiener functional (1.3) is carried out in §6.

Note. The author was informed that Professor N. Ikeda has also obtained results related to the present work.

2. Assumptions and the main theorem. Unfortunately the Euclidean coordinate system is not suitable for our work. We therefore need a little elementary differential geometry. Although we may sometimes discuss the problem under general differentio-geometrical setting, the case where M is the exterior of a smooth, strictly convex body is our primary concern. We will see that various geometrical assumptions we make along the way are satisfied in this important case.

So let us assume that (M,g) is a Riemannian manifold with smooth boundary ∂M. We assume that ∂M is strictly concave when viewed from M. Mathematically this means that the second fundamental form (defined below) of ∂M is strictly positive definite. Now let a and b be two points on ∂M such that there is a unique geodesic in ∂M joining them on which they are not conjugate. For example, a and b can be any two nonantipodal points on a sphere. The geodesic is the arc of the great circle passing through a and b of lesser length. We can set up a semigeodesic coordinate system $\tilde{x} = (x^2, \cdots, x^d)$ on ∂M in a neighborhood of γ with a as the origin and x^2 in the direction of the geodesic γ (cf. Molchanov [9, p. 10]). We let $x = (x^1, \tilde{x})$ be the

point in M which lies on the geodesic passing through $\tilde{x}$ and perpendicular to ∂M with $x^1 = d(x, \partial M)$.

Instead of M, which has a boundary, we can consider $(M^- \cup M, g)$ the double of M. Here M^- is just a copy of M, and M and M^- are identified along the boundary. The heat kernel on $M^- \cup M$ and the Neumann heat kernel on M are related in a very simple way (see §3 below).

The second fundamental form H of the boundary ∂M can be identified with the matrix

$$H_{ij} = H\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \left\langle \nabla_i \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^1} \right\rangle$$

($\nabla_i = \nabla_{\partial/\partial x^i}$ is the covariant derivative). The normal curvature of γ at $\gamma(s)$ is by definition

$$N(s) = H_{\gamma(s)}(\dot{\gamma}(s), \dot{\gamma}(s))$$

(see [7, p. 44]). For brevity, we sometimes write $H_{ij}(s)$ for $H_{ij}(\gamma(s))$ and $g(s)$ for $g(\gamma(s))$. The following lemma clarifies the geometric meaning of the second fundamental form.

LEMMA 2.1. *Let $g = (g_{ij})$ be the metric matrix in the semigeodesic coordinates.* (a) *We have*

$$g_{1i}(x) = \delta_{1i}, \qquad g_{2i}(0, \tilde{x}) = \delta_{2i}, \quad i = 1, \cdots, d.$$

(b) *Near the boundary ∂M, the metric matrix has the expansion*

$$g_{ij}(x) = g_{ij}(0, \tilde{x}) + 2H_{ij}(\tilde{x})|x^1| + O(|x^1|^2), \qquad 2 \le i, j \le d.$$

Proof. Since the coordinate line $\tilde{x} = \text{const.}$ is a geodesic perpendicular to N, we have $g_{1i}(0, \tilde{x}) = \delta_{1i}$ for $x \in \partial M$ and $\nabla_1(\partial/\partial x^1) = 0$. This implies

$$\nabla_1 g_{1i} = \left\langle \frac{\partial}{\partial x^1}, \nabla_1 \frac{\partial}{\partial x^i} \right\rangle = \frac{1}{2}\nabla_i \left\langle \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^1} \right\rangle = 0.$$

It follows that $g_{1i}(x) = \delta_{1i}$. The same proof applies to $g_{2i}(0, \tilde{x})$. Part (a) is proved.

By the definition of the second fundamental form and part (a), we have on ∂M

$$H_{ij}(\tilde{x}) = \left\langle \frac{\partial}{\partial x^j}, \nabla_i \frac{\partial}{\partial x^1} \right\rangle = \left\langle \frac{\partial}{\partial x^i}, \nabla_1 \frac{\partial}{\partial x^j} \right\rangle = \frac{1}{2}\nabla_1 g_{ij}.$$

Part (b) follows immediately.

We will prove our asymptotic formula for the heat kernel under the following two geometrical assumptions.

Assumption (A). The normal curvature $N(s) = H_{22}(s), 0 \le s \le \rho$, is *strictly* positive along the geodesic γ.

Assumption (B). For any neighborhood G of γ in M, there exists $\epsilon > 0$ such that any piecewise smooth curve in M joining a and b with length $\le d(a, b) + \epsilon$ lies completely inside G. Equivalently, $d(a, b) < d(a, \partial G) + d(b, \partial G)$ for any neighborhood G of γ.

It is easy to verify that in the case where M is the exterior of a strictly convex body in the Euclidean space, the above assumptions (A) and (B) are satisfied.

Let (μ_1, ϕ_1) be the first normalized eigenpairs of the eigenvalue problem

$$\frac{1}{2}\phi''(x) - |x|\phi(x) + \mu\phi(x) = 0, \qquad x \in R^1.$$

We are in a position to state our main result.

THEOREM. *Let M be a Riemannian manifold with boundary and $p(t,x,y)$ the heat kernel of the Laplace–Beltrami operator $\Delta/2$ on M under the Neumann boundary condition (insulated boundary). Suppose that a and b are two points on the boundary such that there is a unique geodesic in the boundary ∂M joining them along which they are not conjugate. Then under further assumptions* (A) *and* (B), *we have as $t \to 0$,*

$$p(t,a,b) \approx \gamma H(a,b)\rho^{2/3}[N(a)N(b)]^{1/6}t^{-(d/2+1/6)} \exp\left\{-\frac{\rho^2}{2t} - \frac{\mu_1\rho^{1/3}}{t^{1/3}}\int_0^\rho N(s)^{2/3}ds\right\},$$

where

$$\gamma = 2(2\pi)^{-(d-1)/2}|\phi_1(0)|^2$$

and

$$H(a,b) = \frac{[g(a)g(b)]^{-1/4}}{\left[\det\int_0^\rho g(s)^{-1}ds\right]^{1/2}}\rho^{2/3}.$$

Remark. $H(a,b)$ has an intrinsic geometric meaning, cf. Molchanov [9, p. 14–15].

Before proving this theorem, we need to transform Assumption (B) into a form more suitable for computation. Let G be any neighborhood of γ; by Assumption (B) we have $d(a,b) < d(a,\partial G) + d(b,\partial G)$. One important consequence of this assumption is that the computation of the asymptotic behavior of $p(t,a,b)$ can be localized inside G. This means that the metric outside G has no effect on the asymptotics of $p(t,a,b)$. In fact if p_{g_1} and p_{g_2} are two heat kernels for the metrics g_1 and g_2 which coincide on G, then we have

$$\lim_{t\to 0}\frac{p_{g_1}(t,a,b)}{p_{g_2}(t,a,b)} = 1$$

(See Azencott [1, p. 157]). Note that in [1], the above relation is proved under the assumption $d(a,b) < \max\{d(a,\partial G), d(b,\partial G)\}$. The result holds, however, under the more relaxed condition $d(a,b) < d(a,\partial G)+d(b,\partial G)$. See Hsu [4] for details. Therefore, for the purpose of computing the asymptotics of $p(t,a,b)$, we may arbitrarily alter the metric outside G to facilitate the computation. Thus we can assume that $M = R^n_+ = \{x = (x^1, x^2, \cdots, x^n) : x^1 \geq 0\}$, $M^- \cup M = R^n$, and that the metric is Euclidean outside a small neighborhood G of γ. Let $g^{-1} = (g^{ij})$ be the inverse of the metric matrix. From Lemma 2.1 a simple calculation shows

$$g^{22}(x) = 1 - 2H_{22}(\tilde{x})|x^1| + O(|x^1|^2).$$

We can then impose the following global assumptions on g^{22}:

Assumption (B1). For all $x \in R^n$, we have $g^{22}(x) \leq 1$.

Assumption (B2). There exists a constant $\gamma > 0$ such that for all $x \in R^n$,

$$1 - 2H_{22}(\tilde{x})|x^1| - \gamma|x^1|^2 \leq g^{22}(x) \leq 1 - 2H_{22}(\tilde{x})|x^1| + \gamma|x^1|^2.$$

The reason that we can make Assumptions (B1) and (B2) is simple: These two assumptions hold on a small neighborhood G of the geodesic γ. We can then choose the metric so that they also hold outside G. Let us emphasize once more that (B1) and (B2) are derived from (A) and (B) and the above mentioned localization principle. We may prove our main theorem under these assumptions without losing the generality of our result.

Remark. A casual reader might think Assumption (B) is redundant because it should always hold. (B) may fail if M is not complete in its Riemannian metric. Since the Euclidean space is complete, (B) indeed holds in this case. On the other hand, Assumption (A) is essential.

Finally let us look at a simple example where (B1) and (B2) are satisfied by the obvious choice of coordinates.

Example. Let $M \subset R^2$ be the exterior component of the ellipse: $x = a\cos\theta, y = \sin\theta$. Introduce coordinates (θ, t) on M:

$$x = \left(a + \frac{bt}{\lambda(\theta)}\right)\cos\theta, \qquad y = \left(b + \frac{at}{\lambda(\theta)}\right)\sin\theta$$

where $\lambda(\theta)^2 = a^2\sin^2\theta + b^2\cos^2\theta$. A simple calculation shows

$$dx^2 + dy^2 = dt^2 + \lambda(\theta)^2\left(1 + \frac{ab}{\lambda(\theta)^3}t\right)^2 d\theta^2.$$

Let

$$x^1 = t, \qquad x^2 = \int_0^\theta \lambda(u)du.$$

We have

$$g^{22}(x) = \left(1 + \frac{ab}{\lambda(\theta)^3}x^1\right)^{-2}$$

and

$$H_{22}(x^2) = \frac{ab}{\lambda(\theta)}.$$

Clearly, Assumptions (A), (B), (B1), and (B2) are satisfied.

3. Proof of the theorem. Let $g^{-1} = (g^{ij})$ be the inverse of the metric matrix g. The Laplace–Beltrami operator on $(M^- \cup M = R^d, g)$ is given by

$$\Delta = \frac{1}{\sqrt{\det g}}\frac{\partial}{\partial x^i}\left(\sqrt{\det g}\, g^{ij}\frac{\partial}{\partial x^j}\right) = g^{ij}\frac{\partial^2}{\partial x^i \partial x^j} + 2b^i\frac{\partial}{\partial x^i}$$

where

$$b^i = \frac{1}{2}\frac{1}{\sqrt{\det g}}\frac{\partial}{\partial x^j}(\sqrt{\det g}\, g^{ij}).$$

Let $X = \{X_t = (X_t^1, \cdots, X_t^d) : t \geq 0\}$ be the Riemannian Brownian motion on $(M^- \cup M, g)$, i.e., the diffusion process generated by $\Delta/2$ (cf. Ikeda–Watanabe [6, Chaps. IV, V]). Denote by $\tau : M^- \cup M \to M$ the natural projection. Then the process $\tau(X)$ is the reflecting Brownian motion on (M, g). Let p^X and $p^{\tau(X)} = p$ be the respective transition density. Then obviously

$$p(t, x, y) = p^X(t, x, y) + p^X(t, x, y^*)$$

where $\{y, y^*\} = \tau^{-1}(y)$. In particular, since $b^* = b$

$$p(t,a,b) = 2p^X(t,a,b). \tag{3.1}$$

Diffusion process X can be obtained as the solution of the stochastic differential equation on R^d:

$$dX_s = \sigma(X_s)\,dB_s + b(X_s)\,ds.$$

Here σ is a smooth square root of g and B is a standard Brownian motion in R^d.

The behavior of the heat kernel $p(t,a,b)$ depends on the law of the Brownian bridge from a and conditioned to reach b at time t. As $t \to 0$, the Brownian bridge tends to travel along the geodesic γ with uniform speed ρ/t. Let

$$Y_s = X_s - \frac{s\rho}{t}e_2.$$

(e_2 is the unit vector $(0,1,0,\cdots,0)$ in x^2-direction.) We therefore expect Y to be a process with small magnitude. The equation for Y is

$$dY_s = \sigma\left(Y_s + \frac{s\rho}{t}e_2\right)dB_s + \left[b\left(Y_s + \frac{s\rho}{t}e_2\right) - \frac{\rho}{t}e_2\right]ds.$$

We now alter the drift of this equation by the Girsanov transform. Consider a new equation

$$dZ_s = \sigma\left(Z_s + \frac{s\rho}{t}e_2\right)dB_s + c\left(Z_s + \frac{s\rho}{t}e_2; t\right)ds. \tag{3.2}$$

Let P^Y and P^Z denote the laws of the processes Y and Z on the sample path space $C([0,t] \to R^n)$. By the Girsanov formula (Ikeda–Watanabe [6, p. 180]), we have

$$\frac{dP^Y}{dP^Z} = \exp\left[\int_0^t \left\langle h\left(Z_s + \frac{s\rho}{t}e_2; t\right), dB_s\right\rangle - \frac{1}{2}\int_0^t \left\|h\left(Z_s + \frac{s\rho}{t}e_2; t\right)\right\|^2 ds\right] \stackrel{\text{def}}{=} N_t.$$

where

$$h(x;t) = \sigma(x)^{-1}\left[b(x) - c(x;t) - \frac{\rho}{t}e_2\right]. \tag{3.3}$$

Let D be a neighborhood of a (the origin of the coordinates). We have

$$\begin{aligned} P[X_t \in D + \rho e_2] &= P[Y_t \in D] = E[N_t; Z_t \in D] \\ &= \int_D E\left[N_t \middle| Z_t = y\right] P\left[Z_t \in dy\right]. \end{aligned}$$

It follows from this and (3.2) that

$$p(t,a,b) = \frac{2}{\sqrt{\det g(\rho)}} E[N_t \big| Z_t = 0]\, p^Z(0,0;t,0). \tag{3.4}$$

where $p^Z(s,z;v,x)$ is the transition density of the process Z (with respect to the Lebesgue measure) defined by (3.2). Formula (3.4) is the key to the subsequent discussion.

We now choose the drift c in (3.2):

$$c(x;t) = b(x) - b^1(x)\, e_1 - \frac{\rho}{t}\,[I - g(x)^{-1}]\, e_2. \tag{3.5}$$

Or, what is the same thing (see (3.3))

$$h(x;t) = \sigma(x)\left[b^1(x)\, e_1 - \frac{\rho}{t} e_2\right]. \tag{3.6}$$

The advantage of this choice will be clear later. Note that

$$c^1\left(z + \frac{s\rho}{t} e_2\right) \equiv 0.$$

This means that the first component Z^1 of (3.2) is simply a one-dimensional Brownian motion.

The last two factors on the right-hand side of (3.4) will now be analyzed separately. First of all, we have the following lemma.

LEMMA 3.1. *As $t \to 0$, we have*

$$p^Z(0,0;t,0) = \left(\frac{1}{2\pi t}\right)^{d/2} H_1\left[1 + O(\sqrt{t})\right]$$

where

$$H_1 = \rho^{d/2}\left[\det \int_0^\rho g(s)^{-1} ds\right]^{-1/2}.$$

To study $E[N_t \big| Z_t = 0]$, set

$$Z_s^* = Z_s + \frac{s\rho}{t} e_2$$

for brevity. Using (3.6), we verify easily

$$\begin{aligned}
&\int_0^t \langle h(Z_s^*;t), dB_s\rangle \\
&\qquad = \int_0^t \langle \sigma(Z_s^*)^{-1} h(Z_s^*;t),\, dZ_s - c(Z_s^*;t)\, ds\rangle \\
&\qquad = -\frac{\rho^2}{t^2}\int_0^t [1 - g^{22}(Z_s^*)]\, ds - \frac{\rho}{t} Z_t^2 + \int_0^t b^1(Z_s^*)\, dZ_s^1 + \frac{\rho}{t}\int_0^t b^2(Z_s^*)\, ds.
\end{aligned}$$

We also have

$$\|h(Z_s^*;t)\|^2 = \frac{\rho^2}{t^2} - \frac{\rho^2}{t^2}[1 - g^{22}(Z_s^*)] + |b^1(Z_s^*)|^2.$$

It follows that

$$\log N_t = -\frac{\rho^2}{2t} - \Theta_t + \log H_2 + F_t \tag{3.7}$$

with

$$\Theta_t = \frac{\rho^2}{2\, t^2}\int_0^t [1 - g^{22}(Z_s^*)]\, ds \tag{3.8}$$

$$H_2 = \exp\left\{\int_0^\rho b^2(se_2)\,ds\right\} = \left[\frac{\det g(\rho)}{\det g(0)}\right]^{1/4}$$

and

(3.9)

$$F_t = -\frac{\rho}{t}Z_t^2 + \int_0^t b^1(Z_s^*)\,dZ_s^1 + \frac{\rho}{t}\int_0^t \left[b^2(Z_s^*) - b^2\left(\frac{s\rho}{t}e_2\right)\right] ds - \frac{1}{2}\int_0^t |b^1(Z_s^*)|^2 ds.$$

It is clear now that the proof of the main theorem in §2 will be completed if we show the following three lemmas.

LEMMA 3.2. *Let $\hat{W}$ be the standard Brownian bridge. We have*

$$\lim_{t\to 0} \frac{E\left[\exp\{-\Theta_t\}\middle| Z_t = 0\right]}{E\left[\exp\left\{-\frac{\rho^2}{\sqrt{t}}\int_0^1 N(s\rho)|\hat{W}_s|ds\right\}\right]} = 1.$$

LEMMA 3.3. *We have*

$$\lim_{t\to 0} \frac{E\left[\exp\{-\Theta_t + F_t\}\middle| Z_t = 0\right]}{E\left[\exp\{-\Theta_t\}\middle| Z_t = 0\right]} = 1.$$

Let

$$S(\lambda; l) \overset{\text{def}}{=} E\left[\exp\left\{-\lambda\int_0^1 l(s)|\hat{W}_s|ds\right\}\right].$$

LEMMA 3.4. *Let $l : [0,1] \to R_+$ be twice continuously differentiable and strictly positive on the closed interval $[0,1]$, then for any $k \geq 0$*

$$S(\lambda; l) = \sqrt{2\pi}|\phi_1(0)|^2[l(0)\,l(1)]^{1/6}\lambda^{1/3}\exp\left\{-\mu_1\lambda^{2/3}\int_0^1 l(s)^{2/3}ds\right\}[1 + O(\lambda^{-k})].$$

The next three sections are devoted to the proof of Lemmas 3.1 to 3.4.

4. Proof of Lemma 3.1. Throughout the rest of this paper, letters $c_1, c_2, \cdots$, whose values may change from one appearance to another, represent constants depending only on the geometry of the manifold.

By (3.2), the function $p^Z(s,z;v,y)$ is the fundamental solution of the parabolic operator

$$L = \frac{\partial}{\partial s} + \frac{1}{2}g^{ij}\left(z + \frac{s\rho}{t}e_2\right)\frac{\partial^2}{\partial z^i \partial z^j} + c^i\left(z + \frac{s\rho}{t}e_2; t\right)\frac{\partial}{\partial z^i}. \tag{4.1}$$

Let us investigate the coefficients c more carefully. First of all, as we have pointed out before, $c^1 \equiv 0$. By Lemma 2.1

$$g(x)^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{g}(\tilde{x})^{-1} - 2\tilde{g}(\tilde{x})^{-1}H(\tilde{x})\tilde{g}(\tilde{x})^{-1}|x^1| + O(|x^1|^2) \end{pmatrix}.$$

($\tilde{g}$ is the last $(n-1)\times(n-1)$ principal minor of g). Hence near the geodesic γ

$$\tilde{c}(z + se_2) = \tilde{b}(z + se_2) - \frac{2\rho}{t}D(s)\,e_2|z^1| + \frac{1}{t}O(|z^1|\|z\|) \tag{4.2}$$

where

$$D(s) = \tilde{g}(se_2)^{-1} H(se_2)\, \tilde{g}(se_2)^{-1}.$$

We prove Lemma 3.1 by the method of parametrix (cf. Friedman [3]). We need to pay special attention to the dependence of the coefficients on t.

Let L^x be the operator obtained from L by freezing the coefficients of L at $z = x$. Set for a positive definite matrix A

$$\Gamma(A, y) = \frac{1}{(2\pi)^{d/2}\sqrt{\det A}} \exp\left\{-\frac{1}{2}\langle y, A^{-1}y\rangle\right\}.$$

Let

$$u_0(s, z; v, x) = \Gamma(\lambda_x(s, v), x - z)$$

with

$$\lambda_x(s, v) = \int_s^v g\left(x + \frac{l\rho}{t}e_2\right)^{-1} dl.$$

We have

$$u_0(0, 0; t, 0) = \left(\frac{1}{2\pi t}\right)^{d/2} H_1 \tag{4.3}$$

with the same H_1 as in the statement of Lemma 3.1.

Now $p^Z = u$ can be obtained by iteration from the equation

$$u(s, z; v, x) = u_0(s, z; v, x) + \int_s^v dl \int_{R^d} u(s, z; l, y)(L - L^x)\, u_0(l, y; v, x) dy.$$

We thus obtain an absolutely convergent series $p^Z = \sum_{m=0}^{\infty} u_m$. Using the easy estimate

$$\|g(z) - g(x)\| + t\|c(z; t) - c(x; t)\| \le c_3\|z - x\| \tag{4.4}$$

which follows from (3.5), we verify by induction the following estimate:

$$|u_m(s, z; v, x)| \le \frac{c_1 c_2^m}{\Gamma(m/2 + 1)} \left[1 + \frac{\sqrt{v - s}}{t}\right] (v - s)^{(m-d)/2} \exp\left\{-\frac{\|z - x\|^2}{c_1(v - s)}\right\}.$$

It follows immediately that

$$p^Z(s, z; x, v) \le \frac{c_4}{(v - s)^{d/2}} \left[1 + \frac{v - s}{t}\right] \exp\left\{-\frac{\|z - x\|^2}{c_4(v - s)}\right\}. \tag{4.5}$$

Now that we have

$$\sum_{m=2}^{\infty} |u_m(0, 0; t, 0)| \le c_5 t^{-(d-1)/2},$$

it is easy to see from (4.3) that the assertion of Lemma 3.1 is implied by the inequality

$$|u_1(0, 0; t, 0)| \le c_6 t^{-(d-1)/2}, \tag{4.6}$$

which we are about to show. By the iteration formula

$$u_1(0,0;t,0) = \int_0^t dl \int_{R^d} \Gamma(\lambda_y(0,l),y)[L - L^0]\Gamma(\lambda_0(l,t),y)\,dy. \tag{4.7}$$

From (4.2), we have

$$L - L^0 = \alpha^{ij}\frac{\partial^2}{\partial y^i \partial y^j} + \beta^i \frac{\partial}{\partial y^i} + \frac{\rho}{t}\gamma^i \frac{\partial}{\partial y^i} - \frac{2\rho}{t}\sum_{i=2}^{n} D_{2i}\left(\frac{s\rho}{t}\right)|y^1|\frac{\partial}{\partial y^i}$$

with $\alpha = \alpha(y,l,t)$, etc., satisfying

$$\|\alpha\| + \|\beta\| + \|\gamma\|^{1/2} \le c_7\|y\|. \tag{4.8}$$

This fact together with (4.7) gives

$$\begin{aligned} u_1(0,0;t,0) = &-\frac{2\rho}{t}\int_0^t D_{2i}\left(\frac{l\rho}{t}\right) dl \int_{R^n} [\Gamma(\lambda_y(0,l),y) - \Gamma(\lambda_0(0,l),y)]|y^1| \\ &\times \frac{\partial}{\partial y_i}\Gamma(\lambda_0(l,t),y)dy + O(t^{-(d-1)/2}). \end{aligned} \tag{4.9}$$

(Inserting $\Gamma(\lambda_0(0,l),y)$ creates a term equal to zero after integration.) Finally using the inequality

$$|\Gamma(\lambda_y(0,l),y) - \Gamma(\lambda_0(0,l),y)| \le c_8 l^{-d/2}\|y\|e^{-\|y\|/c_8 l}$$

we obtain (4.6) from (4.9) by simple estimation. The proof of Lemma 3.1 is therefore complete.

5. Proof of Lemma 3.2 and Lemma 3.3. We adopt the following notational convention. If $G(Z)$ is a functional of the process Z, the same functional of Z conditioned by $Z_t = 0$ is denoted by $\hat{G}$, i.e., $\hat{G} = G(\hat{Z})$. Also if $x \in R^n$, then $\tilde{x} = (x^2, \cdots, x^d)$.

Set

$$Z_s^t = \frac{Z_{st}}{\sqrt{t}}, \qquad M^t = \max_{0\le s\le 1}\|Z_s^t\|, \qquad \tilde{M}^t = \max_{0\le s\le 1}\|\tilde{Z}_s^t\|.$$

As mentioned immediately before Lemma 3.1, the process

$$\{W_s \overset{\text{def}}{=} Z_s^{t,1}; 0 \le s \le 1\}$$

is a one-dimensional Brownian motion.

Let P_W be the law of $\tilde{Z}^t$ conditioned by the process W. This means that under the probability P_W, the process $\tilde{Z}^t$ is the solution of the stochastic differential equation

$$d\tilde{Z}_s^t = \tilde{\sigma}(\sqrt{t}W_s e_1 + \sqrt{t}\tilde{Z}_s^t + s\rho e_2)\,d\tilde{B}_s + \sqrt{t}\tilde{c}(\sqrt{t}W_s e_1 + \sqrt{t}\tilde{Z}_s^t + s\rho e_2)\,ds.$$

In this equation $W = \{W_s; 0 \le s \le 1\}$ is assumed to be deterministic. Let p^{Z^t} be the transition density of Z^t and let $p_W^{\tilde{Z}^t}$ be that of the process $\tilde{Z}^t$ under the probability P_W.

Inequality (4.6) and Lemma 3.1 can be paraphrased as follows:

$$p^{Z^t}(s,z;v,x) \le \frac{c_2}{(v-s)^{d/2}} e^{-\|z-x\|^2/c_2(v-s)} \le \frac{c_3}{\|z-x\|^d}$$

$$p^{Z^t}(0,0;1,0) = \left(\frac{1}{2\pi}\right)^{d/2} [1+O(\sqrt{t})] \ge c_4.$$

The proof of Lemma 3.1 can be applied to obtain the following estimates for function $p^{\tilde{Z}^t}_W$:

$$p^{\tilde{Z}^t}_W(s,\tilde{z};v,\tilde{x}) \le \frac{c_2}{(v-s)^{(d-1)/2}} e^{-\|\tilde{z}-\tilde{x}\|^2/c_2(v-s)} \le \frac{c_3}{\|\tilde{z}-\tilde{x}\|^{(d-1)}} \tag{5.1a}$$

and

$$p^{\tilde{Z}^t}_W(0,0;t,0) \ge c_4 \left[1 - c_5 \int_0^1 |W_s| ds\right]. \tag{5.1b}$$

To see this, we only need to observe that (4.2) and estimates (4.4) and (4.8), which are crucial to the proof there, should be replaced by

$$\tilde{c}(\sqrt{t}W_{s/t}e_1 + \tilde{z} + se_2) = \tilde{b}(\sqrt{t}W_{s/t}e_1 + \tilde{z} + se_2) - \frac{2\rho}{t}\left[\tilde{I} - \tilde{g}(\sqrt{t}W_{s/t}e_1 + \tilde{z} + se_2)^{-1}\right] e_2 + \frac{1}{\sqrt{t}} O(|W_{s/t}|\,\|\tilde{z}\|)$$

$$\left\|\tilde{g}(\sqrt{t}W_{s/t}e_1 + \tilde{z}) - \tilde{g}(\sqrt{t}W_{s/t}e_1 + \tilde{x})\right\| \le c_6\|\tilde{z}-\tilde{x}\|$$

$$\left\|\tilde{c}(\sqrt{t}W_{s/t}e_1 + \tilde{z};t) - \tilde{c}(\sqrt{t}W_{s/t}e_1 + \tilde{x};t)\right\| \le \frac{c_6}{t}\|\tilde{z}-\tilde{x}\|$$

and

$$\|\alpha\| + \|\beta\| \le c_7\|\tilde{y}\|, \qquad \|\gamma\| \le c_7\sqrt{t}|W_{s/t}|\,\|\tilde{y}\|.$$

with constants c_6, c_7 independent of W and t.

LEMMA 5.1. *There exist constants c_0, c_1 independent of t such that for sufficiently large $a \gg 1$,*

$$P_W\left[\hat{\tilde{M}}^t > a\right] \le e^{-c_1 a^2} \qquad \text{if } \int_0^1 |W_s| ds \le c_0. \tag{a}$$

and

$$P\left[\hat{M}^t > a\right] \le e^{-c_1 a^2}. \tag{b}$$

Proof. Let

$$\tau_a = \inf\left\{s : \|\tilde{Z}^t_s\| \ge a\right\}.$$

By the Markov property, we have, for any neighborhood D of the origin in $R^{(d-1)}$,

$$P_W\left[\tilde{M}^t > a, \tilde{Z}^t_1 \in D\right] = E_W\left[\int_D p^{\tilde{Z}^t}_W(a,\tau_a;1,y)\,dy; \tau_a < 1\right]. \tag{5.2}$$

Divide (5.2) by $P\left[\tilde{Z}_1^t \in D\right]$ and use (5.1). Letting $|D| \to 0$, we see that for $a \gg 1$,

$$P_W\left[\hat{\tilde{M}}^t > a\right] \leq \frac{1}{2}\left[1 - c_5\int_0^1 |W_s|ds\right]^{-1} P_W\left[\tilde{M}^t > a\right]. \tag{5.3}$$

To estimate the last probability, we note that the equation of $\tilde{Z}^t$ is

$$d\tilde{Z}_s^t = dQ_s + \sqrt{t}\tilde{c}(\sqrt{t}W_s e_1 + \sqrt{t}\tilde{Z}_s^t + s\rho e_2)\,ds \tag{5.4}$$

where

$$dQ_s = \tilde{\sigma}(\sqrt{t}W_s e_1 + \sqrt{t}\tilde{Z}_s^t + s\rho e_2)\,d\tilde{B}_s.$$

By Lemma 2.1 and (3.5) the drift in (5.4) is bounded by

$$\sqrt{t}\left\|c(\sqrt{t}W_s e_1 + \sqrt{t}\tilde{Z}_s^t + s\rho e_2; t)\right\| \leq c_8\sqrt{t} + c_8|W_s|.$$

Also note that $Q_s^i, i = 2, \cdots$ are martingale with bounded characteristic: $[Q^i]_1 \leq d\|\sigma\|_\infty^2$. It follows that for some $a \geq 2c_8$ and all $t \leq 1$

$$\begin{aligned} P_W\left[\tilde{M}^t > a\right] \leq & P_W\left[\max_{0\leq s\leq 1}\|Q_s\| > \frac{a}{2}\left(1 - \int_0^1 |W_s|ds\right)\right] \\ \leq & dP_W\left[\max_{0\leq s\leq d\|\sigma\|_\infty^2}\|\beta_s\| > \frac{a}{2}\left(1 - \int_0^1 |W_s|ds\right)\right] \\ \leq & d\exp\left\{-c_9 a^2\left(1 - \int_0^1 |W_s|ds\right)^2\right\} \end{aligned}$$

(β is an independent one-dimensional Brownian motion). Part (a) follows immediately from (5.3) by choosing, for example, $c_0 < \min(c_5^{-1}, 1)$ and $c_1 > c_9(1 - c_0)^2$. The proof of part (b) is similar and easier.

LEMMA 5.2. *For any positive ϵ, K and $0 < \delta < 1/6$, there exists a positive constant $t_0 = t_0(\epsilon, K, \delta)$ such that for all $t \leq t_0$,*

$$P\left[\int_0^1 |\hat{W}_s|ds \leq Kt^{1/6}, \quad \max_{0\leq s\leq 1}|\hat{W}_s| \geq \epsilon t^{-1/6}\right] \leq \exp\left\{-t^{-(1/3+\delta)}\right\}.$$

Proof. This lemma is proved in Lemma 5.4 of Ikeda [5, p. 188–189].

We now turn to the following proof.

Proof of Lemma 3.2. Set

$$A_{t,K} = \left\{\omega : \int_0^1 |\hat{W}_s|ds \leq Kt^{1/6}\right\}$$

$$B_{t,\epsilon} = \left\{\omega : \max_{0\leq s\leq 1}|\hat{W}_s| \geq \epsilon t^{-1/6}\right\}.$$

Also set

$$G_t = \hat{\tilde{M}}^t \int_0^1 |\hat{W}_s|ds.$$

We have from (3.8)

$$\hat{\Theta}_t = \frac{\rho^2}{2t}\int_0^1 [1 - g^{22}(\sqrt{t}\hat{Z}_s^t + s\rho e_2)]\,ds.$$

Assumption (B2) implies

$$2H_{22}(\tilde{z}+se_2)|z^1| - \gamma|z^1|^2 \le 1 - g^{22}(z+se_2) \le 2H_{22}(\tilde{z}+se_2)|z^1| + \gamma|z^1|^2.$$

Since

$$|H_{22}(\tilde{z}+se_2) - N(s)| \le c_1\|\tilde{z}\|$$

and $N(s)$ is strictly positive by Assumption (A), we have

$$\hat{\Theta}_t \le \frac{\rho^2}{\sqrt{t}}\left[1 + c_4\epsilon\gamma t^{1/3}\right]\int_0^1 N(s\rho)|\hat{W}_s|ds + c_3G_t \text{ on } B_{t,\epsilon}^c. \tag{5.5}$$

Symmetrically, we have

$$\hat{\Theta}_t \ge \frac{\rho^2}{\sqrt{t}}\left[1 - c_4\epsilon\gamma\|_\infty t^{1/3}\right]\int_0^1 N(s\rho)|\hat{W}_s|ds - c_3G_t \quad \text{on } B_{t,\epsilon}^c. \tag{5.6}$$

Now Lemma 5.1(a) implies that if $t \le (c_0/K)^{1/6}$

$$E_{\hat{W}}\left[\exp\{c_3G_t\}\right] \le \exp\left\{c_5\int_0^1 |\hat{W}_s|ds + c_5\left(\int_0^1 |\hat{W}_s|ds\right)^2\right\} \quad \text{on } A_{t,K}. \tag{5.7}$$

(Integration by parts!) By the Schwartz inequality, (5.7) gives

$$E_{\hat{W}}\left[\exp\{-c_3G_t\}\right] \ge \exp\left\{-c_5\int_0^1 |\hat{W}_s|ds - c_5\left(\int_0^1 |\hat{W}_s|ds\right)^2\right\} \quad \text{on } A_{t,K}. \tag{5.8}$$

We also have

$$\hat{\Theta}_t \ge 2c_6Kt^{-1/3} - \frac{\rho^2\gamma}{2}\int_0^1 |\hat{W}_s|^2ds \quad \text{on } A_{t,K}^c.$$

Hence for $\eta = [2c_6/\rho^2\|f\|_\infty]^{1/2}$, we have

$$\hat{\Theta}_t \ge c_6Kt^{-1/3} \quad \text{on } A_{t,K}^c \cap B_{t,\eta\sqrt{K}}^c. \tag{5.9}$$

Observe that

$$E\left[G(\hat{Z})\right] = E\left[E_{\hat{W}}\left[G(\hat{W}e_1 + \hat{\tilde{Z}})\right]\right].$$

Thus, on the one hand, using (5.5), (5.8), and Lemma 5.2, we have

$$\begin{aligned}
&E\left[\exp\left\{-\hat{\Theta}_t\right\}\right] \\
&\quad \ge E\left[\exp\left\{-\hat{\Theta}_t\right\}; A_{t,K}\cap B_{t,\epsilon}^c\right] \\
&\quad \ge E\left[\exp\left\{-\frac{\rho^2}{\sqrt{t}}[1 + c_4\epsilon\gamma t^{1/3} + c_7t^{1/2} + c_7Kt^{2/3}]\int_0^1 N(s\rho)|\hat{W}_s|ds\right\}; A_{t,K}\cap B_{t,\epsilon}^c\right] \\
&\quad \ge E\left[\exp\{\cdots\}\right] - E\left[\exp\{\cdots\}; A_{t,K}^c\right] - P\left[A_{t,K}\cap B_{t,\epsilon}\right] \\
&\quad \ge \left[\exp\left\{-\frac{\rho^2}{\sqrt{t}}[1 + c_4\epsilon\gamma t^{1/3} + c_7t^{1/2} + c_7Kt^{2/3}]\int_0^1 N(s\rho)|\hat{W}_s|ds\right\}\right] \\
&\quad - \exp\left\{-c_8Kt^{-1/3}\right\} - \exp\left\{-c_8t^{-(1/3+\delta)}\right\}.
\end{aligned}$$

On the other hand, by Assumption (B1), we have $\hat{\Theta}_t \geq 0$; hence using (5.6), (5.7), (5.9), and Lemma 5.2, we have

$$\begin{aligned} E\left[\exp\left\{-\hat{\Theta}_t\right\}\right] \leq & E\left[\exp\left\{-\hat{\Theta}_t\right\}; A_{t,K} \cap B_{t,\epsilon}^c\right] + E\left[\exp\left\{-\hat{\Theta}_t\right\}; A_{t,K}^c \cap B_{t,\eta\sqrt{K}}^c\right] \\ & + P\left[B_{t,\eta\sqrt{K}}\right] + P[A_{t,K} \cap B_{t,\epsilon}] \\ \leq & E\left[\exp\left\{-\frac{\rho^2}{\sqrt{t}}[1 - c_4\epsilon\gamma t^{1/3} - c_7 t^{1/2} - c_7 K t^{2/3}] \int_0^1 N(s\rho)|W_s| ds\right\}\right] \\ & + \exp\left\{c_6 K t^{-1/3}\right\} + \exp\left\{-c_8 K t^{-1/3}\right\} + \exp\left\{-t^{-(1/3+\delta)}\right\}. \end{aligned}$$

By Lemma 3.4, which we will prove independently in §6, there exist constants c_9 and c_{10} such that as $t \to 0$, for any $k \geq 0$,

$$E\left[\exp\left\{-\frac{\rho^2}{\sqrt{t}} \int_0^1 N(s\rho)|\hat{W}_s| ds\right\}\right] \sim c_9\, t^{-1/6} \exp\left\{-c_{10} t^{-1/3}\right\}. \tag{5.10}$$

Choose $K > c_{10}/\min(c_6, c_8)$. Using (5.10) and the above bounds for $E\left[\exp\{-\hat{\Theta}_t\}\right]$, we obtain

$$\exp\{-c_{11}\epsilon\gamma\} \leq \lim_{t\to 0}\left\{\begin{matrix}\sup \\ \inf\end{matrix}\right\} \frac{E\left[\exp\{-\hat{\Theta}_t\}\right]}{E\left[\exp\left\{-\frac{\rho^2}{\sqrt{t}} \int_0^1 N(s\rho)|\hat{W}_s| ds\right\}\right]} \leq \exp\{c_{11}\epsilon\gamma\}.$$

($c_{11} = 2\, c_4 c_{10}/3$.) Letting $\epsilon \to 0$, we obtain Lemma 3.2.

Proof of Lemma 3.3. Let us first prove: There exist constants c_1 and c_2 such that for any $|q| \geq 1$

$$E\left[\exp\{q\hat{F}_t\}\right] \leq c_1 e^{c_2 q^2 t}. \tag{5.11}$$

Set

$$\begin{aligned} C_u = & \sqrt{t} \int_0^u b^1(\sqrt{t}\hat{Z}_s^t + s\rho e_2)\, d\hat{W}_s + \int_0^u \left[b^2(\sqrt{t}\hat{Z}_s^t + s\rho e_2) - b^2(s\rho e_2)\right] ds \\ & + \int_0^u \left|b^1(\sqrt{t}\hat{Z}_s^t + s\rho e_2)\right|^2 ds. \end{aligned}$$

Then we have $\hat{F}_t = C_1$. Obviously

$$\left\{E\left[\exp\{q\hat{F}_t\}\right]\right\}^2 \leq E\left[\exp\{2qC_{1/2}\}\right] E\left[\exp\{2q(C_1 - C_{1/2})\}\right].$$

Thus it is enough to prove the estimate for each of the factors on the right-hand side of the above inequality. The proofs for the two factors are the same. Take, for example, the first factor. Since $\hat{W}$ is a standard Brownian bridge, we can write

$$d\hat{W}_s = dW_s - \frac{\hat{W}_s}{1-s} ds$$

for a Brownian motion W. Set

$$A_u = 2q\sqrt{t}\int_0^u b^1(\sqrt{t}\hat{Z}_s^t + s\rho e_2)\,dW_s - 2q^2t\int_0^u \left|b^1(\sqrt{t}\hat{Z}_s^t + s\rho e_2)\right|^2 ds$$

and

$$D_u = 2qC_u - A_u.$$

We have

$$\{E\left[\exp\{2qC_{1/2}\}\right]\}^2 \le E\left[\exp\{2A_{1/2}\}\right] E\left[\exp\{2D_{1/2}\}\right]. \tag{5.12}$$

The first factor on the right-hand side is equal to 1 by the choice of A_u. As for the second factor, we have the bound

$$2|D_{1/2}| \le c_3q^2t + c_3q\sqrt{t}M^t.$$

It follows immediately from Lemma 5.1(b) that the second factor in (5.12) is bounded by $c_4e^{2c_2q^2t}$. This implies

$$E\left[\exp\{qC_{1/2}\}\right] \le c_1e^{c_2q^2t}.$$

Inequality (5.11) is proved.

We now complete the proof of Lemma 3.3. By Lemma 3.2, we have

$$E\left[\exp\left\{-c\hat{\Theta}_t\right\}\right] \ge \exp\left\{-c_5^{-1}t^{-1/3}\right\} \tag{5.13}$$

for fixed c. We use the cases $c = 1, 2$. Let $p > 1$ and $1/p + 1/q = 1$. By (5.11), (5.13) and the Schwartz inequality we obtain

$$\begin{aligned}
&E\left[\exp\left\{-\hat{\Theta}_t + \hat{F}_t\right\}\right] \\
&\quad\le \left\{E\left[\exp\left\{-\hat{\Theta}_t\right\}\right]\right\}^{1/p}\left\{E\left[\exp\left\{-2\hat{\Theta}_t\right\}\right]\right\}^{1/2q}\left\{E\left[\exp\left\{2q\hat{F}_t\right\}\right]\right\}^{1/2q} \\
&\quad\le E\left[\exp\left\{-\hat{\Theta}_t\right\}\right] c_1^{1/2q}\exp\left\{\frac{c_5}{2}(q^3t)^{-1/3} + 2c_2qt\right\}
\end{aligned}$$

and

$$\begin{aligned}
&E\left[\exp\left\{-\hat{\Theta}_t + \hat{F}_t\right\}\right] \\
&\quad\ge \left\{E\left[\exp\left\{-\hat{\Theta}_t\right\}\right]\right\}^{p}\left\{E\left[\exp\left\{-2\hat{\Theta}_t\right\}\right]\right\}^{-p/2q}\left\{E\left[\exp\left\{-\frac{2q}{p}\hat{F}_t\right\}\right]\right\}^{-p/2q} \\
&\quad\ge E\left[\exp\left\{-\hat{\Theta}_t\right\}\right] c_1^{-p/2q}\exp\left\{-\frac{c_5}{2}p(q^3t)^{-1/3} - 2c_2p^{-1}qt\right\}.
\end{aligned}$$

Taking $q = t^{-1/2}$ and letting $t \to 0$, we obtain immediately Lemma 3.3.

6. Proof of Lemma 3.4. Assume first that l is a constant. Using the scaling property of Brownian motion, we have

$$S(\lambda; t) = E\left[\exp\left\{-\int_0^{(\lambda l)^{2/3}} |W_s|ds\right\}\middle| W_{(\lambda l)^{2/3}} = 0\right]$$

(W is one-dimensional Brownian motion). By the Feynmann–Kac formula,

$$S(\lambda; l) = \sqrt{2\pi}(\lambda l)^{1/3} q((\lambda l)^{2/3}, 0, 0)$$

where $q(s, x, y)$ is the fundamental solution of

$$L = \frac{\partial}{\partial t} - \frac{1}{2}\frac{\partial^2}{\partial x^2} + |x|.$$

We have the eigenexpansion

$$q(s, x, y) = \sum_{m=0}^{\infty} \exp\{-\mu_m s\}\phi_m(x)\phi_m(y).$$

It follows easily that for any $k \geq 0$,

$$S(\lambda; l) = \sqrt{2\pi}|\phi_1(0)|^2(\lambda l)^{1/3} \exp\left\{-\mu_1(\lambda l)^{2/3}\right\}\left[1 + O(\lambda^{-k})\right]. \tag{6.1}$$

Thus Lemma 3.4 holds in this special case. For the general case, let

$$\begin{aligned}
L(s) &= \int_0^s l(u)^{2/3} du \\
\mu(s) &= \frac{1}{3}\frac{d}{ds}\left[\log l(L^{-1}(s))\right] \\
\psi(s) &= L(1)\,\mu(L(1)\,s) \\
T_\lambda &= \lambda^{2/3} L(1) \\
J_\psi &= \frac{1}{2}\int_0^1 [\psi'(s) + \psi(s)^2]|\hat{W}_s|^2 ds \\
\Omega(\lambda, \psi) &= J_\psi + \lambda \int_0^1 |\hat{W}_s| ds
\end{aligned}$$

(L^{-1} is the inverse function of L). It is not difficult to see that Lemma 3.4 is implied by the following two relations:

$$S(\lambda; l) = \left[\frac{l(0)\,l(1)}{L(1)^3}\right]^{1/6} E\left[\exp\left\{-\Omega(T_\lambda^{3/2}, \psi)\right\}\right] \tag{6.2}$$

$$\lim_{\lambda \to \infty} \frac{E\left[\exp\{-\Omega(\lambda, \psi)\}\right]}{E\left[\exp\{-\Omega(\lambda, 0)\}\right]} = 1. \tag{6.3}$$

To show (6.2), let δ be the δ-function at $x = 0$, and set

$$u_\lambda(s, x) = E_x\left[\exp\left\{-\lambda \int_s^1 l(s)|W_s| ds\right\} \delta(W_1)\right].$$

We have $S(\lambda; l) = \sqrt{2\pi} u_\lambda(0, 0)$. Function $u_\lambda(s, x)$ satisfies

$$\frac{\partial u}{\partial s} + \frac{1}{2}\frac{\partial^2 u}{\partial x^2} - \lambda l(s)|x|u = 0, \qquad u(1, \cdot) = \delta.$$

Introduce a new function $w_\lambda(s,x)$ by

$$u_\lambda(s,x) = [\lambda l(1)]^{1/3} w_\lambda\left(\lambda^{2/3}L(s), [\lambda l(s)]^{1/3}x\right).$$

Then

$$S(\lambda;l) = \sqrt{2\pi}[\lambda l(1)]^{1/3} w_\lambda(0,0). \tag{6.4}$$

We verify directly that $w_\lambda(s,x)$ satisfies the equation

$$\frac{\partial w}{\partial s} + \frac{1}{2}\frac{\partial^2 w}{\partial x^2} + \lambda^{-2/3}\mu(\lambda^{-2/3}s)\,x\,\frac{\partial w}{\partial x} - |x|w = 0, \qquad w(T_\lambda,\cdot) = \delta$$

By the Girsanov formula and the Feynman–Kac formula, we can write

$$w_\lambda(0,0) = E\left[\exp\{A(\lambda)\}\delta(W_{T_\lambda})\right]$$

where

$$A(\lambda) = \lambda^{-2/3}\int_0^{T_\lambda} \mu\left(\frac{s}{\lambda^{2/3}}\right) W_s dW s - \frac{1}{2}\lambda^{-4/3}\int_0^{T_\lambda} \mu\left(\frac{s}{\lambda^{2/3}}\right)^2 |W_s|^2 ds - \int_0^{T_\lambda} |W_s| ds.$$

Using the scaling property of Brownian motion, we can write

$$w_\lambda(0,0) = \frac{1}{\sqrt{2\pi T_\lambda}} E\left[\exp\{B(\lambda)\}\right] \tag{6.5}$$

with

$$B(\lambda) = \int_0^1 \psi(s)\hat{W}_s d\hat{W}_s - \frac{1}{2}\int_0^1 \psi(s)^2|\hat{W}_s|^2 ds - T_\lambda^{3/2}\int_0^1 |\hat{W}_s| ds.$$

By Itô's formula,

$$\begin{aligned} B(\lambda) &= -\frac{1}{2}\int_0^1 [\psi'(s) + \psi(s)^2]|\hat{W}_s|^2 ds - T_\lambda^{3/2}\int_0^1 |\hat{W}_s| ds - \frac{1}{2}\int_0^1 \psi(s)\,ds \\ &= -\Omega(T_\lambda,\psi) - \frac{1}{6}\log\frac{l(1)}{l(0)}. \end{aligned}$$

The desired formula (6.2) follows from (6.4) and (6.5).

It remains to prove (6.3). We claim

$$\exists p > 1: \qquad C(p) \stackrel{\text{def}}{=} E\left[\exp\{-pJ_\psi\}\right] < \infty. \tag{6.6}$$

Let $\{X_1, X_2, \cdots\}$ be a sequence of independently and identically distributed random variables with standard normal distribution $N(0,1)$. Then the Brownian bridge $\hat{W}_s$ can be expanded as

$$\hat{W}_s = \frac{\sqrt{2}}{\pi}\sum_{k=1}^{\infty} X_k \frac{\sin k\pi s}{k}.$$

We have

$$J_\psi = \frac{1}{2}\sum_{k,l=1}^{\infty} a_{kl} X_k X_l$$

with

$$a_{kl} = \frac{2}{\pi^2}\int_0^1 [\psi'(s) + \psi(s)^2]\frac{\sin k\pi s}{k}\frac{\sin l\pi s}{l}ds.$$

Let H be the Hilbert space

$$H = \left\{f \in AC[0,1]: \quad f(0) = f(1) = 0, \quad \|f\|_H^2 \stackrel{\text{def}}{=} \int_0^1 |f'(s)|^2 ds < \infty\right\}.$$

Let $\{e_1, e_2, \cdots\}$ be an orthonormal basis for H. Define $A : H \to H$ by $Ae_k = \sum_{l=1}^{\infty} a_{kl}e_l$. Let $\alpha_1, \alpha_2, \cdots$ ($\alpha_i \to 0$) be the eigenvalues of A with normalized eigenvectors $f_1, f_2, \cdots$. Define (c_{kl}) by $e_k = \sum_{l=1}^{\infty} c_{kl}f_l$. The random variables $Y_i = \sum_{l=1}^{\infty} c_{li}X_l, i = 1, 2, \cdots$ are again i.i.d. with standard normal $N(0,1)$. Furthermore $J_\psi = \frac{1}{2}\sum_{i=1}^{\infty} \alpha_i |Y_i|^2$. It follows that as long as $1 + p\alpha_i \geq 0$ for all i, we have

$$C(p) = E\left[\exp\left\{-\frac{p}{2}\sum_{i=1}^{\infty}\alpha_i|Y_i|^2\right\}\right] = \prod_{i=1}^{\infty}(1 + p\alpha_i)^{-1/2}. \tag{6.7}$$

The infinite product (6.7) converges to a finite value if and only if the series $\sum_{i=1}^{\infty}\alpha_i$ converges and $1 + p\alpha_i > 0$ for all i. On the other hand, from the definition of $C(p)$, we know $C(p)$ is finite for small p. Thus the the series $\sum_{i=1}^{\infty}\alpha_i$ indeed converges. It is now clear that $C(p)$ is finite for those p such that $1 + p\alpha_i > 0$ for all i. Thus $C(p) < \infty$ for some $p > 1$ if and only if all eigenvalues $\alpha_i > -1$ (note that $\alpha_i \to 0$), or what is the same thing,

$$\forall f \in H: \qquad \langle Af, f\rangle_H > -\|f\|_H^2. \tag{6.8}$$

A direct computation shows that

$$\langle Af, f\rangle = \int_0^1 [\psi'(s) + \psi(s)^2]\,|f(s)|^2 ds.$$

Relation (6.8) follows then from the elementary fact: for all $f \in H$

$$\int_0^1 [\psi'(s) + \psi(s)^2]|f(s)|^2 ds + \int_0^1 |f'(s)|^2 ds = \int_0^1 [\psi(s)f(s) - f'(s)]^2 ds > 0.$$

Equation (6.6) is proved.

We can now finish the proof of (6.3). Let

$$C_{\lambda,\epsilon} = \left\{\omega : |J_\psi| \leq \epsilon\lambda^{1/3}\int_0^1 |\hat{W}_s|ds\right\}$$

and

$$D_{\lambda,K} = \left\{\omega : \int_0^1 |\hat{W}_s|ds > K\lambda^{-1/3}\right\}.$$

Then on $C_{\lambda,\epsilon}$

$$\lambda[1 - \epsilon\lambda^{-2/3}]\int_0^1 |\hat{W}_s|ds \leq \Omega(\lambda, \psi) \geq \lambda[1 + \epsilon\lambda^{-2/3}]\int_0^1 |\hat{W}_s|ds.$$

On the set $D_{\lambda,K}$

$$\Omega(\lambda,\psi) \geq J_\psi + K\lambda^{2/3}.$$

It follows that on the one hand

$$\begin{aligned} E\left[\exp\{-\Omega(\lambda,\psi)\}\right] \leq & E\left[\exp\left\{-\lambda[1-\epsilon\lambda^{-2/3}]\int_0^1 |\hat{W}_s|ds\right\}\right] \\ & + C(1)\exp\{-K\lambda^{2/3}\} + C(p)^{1/p}P\left[C^c_{\lambda,\epsilon}\cap D^c_{\lambda,K}\right]^{1/q}. \end{aligned} \tag{6.9}$$

Note that we have proved $C(1)$ and $C(p)$ are finite. On the other hand, we have

$$\begin{aligned} E\left[\exp\{-\Omega(\lambda,\psi)\}\right] \geq & E\left[\exp\left\{-\lambda[1+\epsilon\lambda^{-2/3}]\int_0^1 |\hat{W}_s|ds\right\}\right] \\ & - \exp\{-K\lambda^{2/3}\} - P\left[C^c_{\lambda,\epsilon}\cap D^c_{\lambda,K}\right]. \end{aligned} \tag{6.10}$$

Note that

$$C^c_{\lambda,\epsilon}\cap D^c_{\lambda,K} \subset \left\{\omega : \int_0^1 |\hat{W}_s|ds \leq K\lambda^{-1/3}, \max_{0\leq s\leq 1}|\hat{W}_s| \geq \epsilon c(\psi)\,\lambda^{1/3}\right\}.$$

with $c(\psi) = \|\psi' + \psi^2\|_\infty^{-1}$. Take $K > \mu_1 L(1)$. By Lemma 5.2 and (6.1), (6.9), and (6.10),

$$e^{-2\mu_1\epsilon/3} \leq \lim_{\lambda\to\infty}\left\{\begin{matrix}\sup\\ \inf\end{matrix}\right\}\frac{E\left[\exp\{-\Omega(\lambda,\psi)\}\right]}{E\left[\exp\{-\Omega(\lambda,0)\}\right]} \leq e^{2\mu_1\epsilon/3}.$$

Letting $\epsilon \to 0$ we obtain (6.3). The proof is complete.

Acknowledgment. The author is grateful to Professor S.R.S. Varadhan for helpful discussions, especially on the material in §6.

REFERENCES

[1] R. AZENCOTT ET AL., *Géodesiques et diffusions en temps petit*, Astérisque, 84–85 (1981), Soc. Math. France.

[2] V. S. BUSLAEV, *Continuum integrals and the asymptotic behavior of the solutions of parabolic equations as $t \to 0$, Applications to diffraction*, Topics in Math. Physics, 2(1968), M. Sh. Birman, ed.

[3] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[4] P. HSU, *Heat kernel on noncomplete Riemannian manifolds*, Annals of Prob., to appear.

[5] N. IKEDA, *On the behavior of the fundamental solutions of the heat equation on certain manifolds*, Taniguchi Symposium on Stochastic Analysis, Katata (1982), pp. 169–195.

[6] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland and Kodansha, 1981.

[7] W. KLINGENBERG, *A Course in Differential Geometry*, Springer-Verlag, New York, Heidelberg, Berlin, 1978.

[8] R. B. MELROSE AND M.E. TAYLOR, *Near peak scattering and the corrected Kirchhoff approximation for a convex obstacle*, Adv. in Math., 55(1985), pp. 242–315.

[9] S. A. MOLCHANOV, *Diffusion processes and Riemannian geometry*, Russian Math. Surveys, 30 (1975), pp. 1–63.

[10] S. R. S. VARADHAN, *Diffusion processes in a small time interval*, Comm. Pure Appl. Math. XX(1967), pp. 659–685.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1128-1144, September 1989

007

GLOBAL EXISTENCE FOR SEMILINEAR PARABOLIC SYSTEMS*

JEFF MORGAN†

Abstract. Global existence results are obtained for semilinear parabolic systems of partial differential equations of the form

$$u_t = D\Delta u + f(u) \quad \text{on } \Omega \times (0, T),$$

with bounded initial data and various boundary conditions, where D is an $m \times m$ diagonal matrix with positive entries on the diagonal, Ω is a smooth bounded domain in R^n, and $f: R^m \to R^m$ is locally Lipschitz. These results are based on f satisfying a Lyapunov-type condition, and generalize a previous result of Hollis, Martin, and Pierre [*SIAM J. Math. Anal.*, 18 (1987), pp. 744-761]. This theory is applied to some specific reaction-diffusion and nerve conduction problems.

Key words. reaction-diffusion systems, global existence, Lyapunov function

AMS(MOS) subject classifications. 35K45, 35B35

1. Introduction and motivation. In recent years, there has been a great deal of research concerning global existence for solutions of semilinear parabolic systems of partial differential equations. The problem has been approached via invariant sets, differential inequalities, semigroup theory, and many other methods (see [1], [3], [16], [19]). More recently, Hollis, Martin, and Pierre [10] studied this problem by examining systems of two unknowns, where one of the unknowns is bounded and the nonlinearity obeys a simple Lyapunov-type condition. The results obtained in this work are a generalization of this idea.

Much of the motivation for this work comes from studying systems of ordinary differential equations. Suppose that m is a positive integer, $f: \mathbf{R}^m \to \mathbf{R}^m$ is continuously differentiable, and $y_0 \in \mathbf{R}^m$. Consider the following system:

$$\begin{aligned} y' &= f(y), \qquad t > 0, \\ y &= y_0, \qquad t = 0. \end{aligned} \tag{1.1}$$

Local existence is well known for such systems. That is, there exists a real number $T_{\max} > 0$ such that (1.1) has a unique noncontinuable solution $y(t)$ on $[0, T_{\max})$. Furthermore, if $T_{\max} < \infty$, then $|y(t)| \to \infty$ as $t \to T_{\max}^-$. Hence, (1.1) has a solution for all $t > 0$ (a global solution) if y does not blow up in finite time. One method for determining whether (1.1) has a solution for all $t > 0$ is the following Lyapunov-type criterion.

PROPOSITION 1.1. *Suppose that $H: \mathbf{R}^m \to [0, \infty)$ is a smooth function satisfying*

(i) $|H(z)| \to \infty$ *as* $|z| \to \infty$;

(ii) *There exists a real number M such that*

$$\nabla H(z) \cdot f(z) \leqq MH(z) \quad \textit{for all } z \in \mathbf{R}^m.$$

Then $T_{\max} = \infty$.

It is interesting to note that if $M \leqq 0$ then the level sets of H determine invariant regions for solutions of (1.1).

Actually, systems of the form (1.1) are a special case of the following. Suppose Ω is a bounded domain in $\mathbf{R}^n$ with a smooth boundary $\partial\Omega$ (say, $\partial\Omega$ is an $(n-1)$-

* Received by the editors September 28, 1987; accepted for publication (in revised form) October 25, 1988.
† Department of Mathematics, Texas A&M University, College Station, Texas 77843.

dimensional $C^{2+\mu}$ manifold, such that Ω lies locally on one side of $\partial\Omega$), D is an $m \times m$ diagonal matrix with positive entries d_i on the diagonal, Δ is the Laplacian operator, ∇ is the gradient operator, ∂ is the derivative operator, $\partial/\partial\eta$ is the derivative with respect to the outward unit normal on $\partial\Omega$, and $v_0 \in L^\infty(\Omega, \mathbf{R}^m)$. Consider the system

$$
\begin{aligned}
&v_t(x,t) = D\Delta v(x,t) + f(v(x,t)), && x \in \Omega, \quad t > 0,\\
&\partial v(x,t)/\partial\eta = 0, && x \in \partial\Omega, \quad t > 0,\\
&v(x,0) = v_0(x), && x \in \Omega.
\end{aligned} \tag{1.2}
$$

Note that if $v_0 \equiv y_0$, then $v(x,t) \equiv y(t)$, where y solves (1.1).

Local existence for (1.2) is similar to that for (1.1) (complete statements are given in § 2). That is, there exists a real number $T_{\max} > 0$ such that (1.2) has a unique, classical, noncontinuable solution $v(x,t) = (v_i(x,t))$ on $\Omega \times [0, T_{\max})$. Furthermore, if $T_{\max} < \infty$, then $|v_i(\cdot,t)|_{\infty,\Omega} \to \infty$ as $t \to T_{\max}^-$, for some $1 \leqq i \leqq m$.

Note that the criterion for determining global existence for (1.2) is essentially the same as that for (1.1). That is, (1.2) has global existence if v does not blow up in finite time in the sup-norm. However, determining global existence for (1.2), even when f is quite simple, can be a difficult task. In general it is not known whether global existence for (1.1) guarantees global existence for (1.2). We might hope that the invariant regions for (1.1) (obtained, for example, when $M = 0$ in Proposition 1.1) would also be invariant regions for (1.2). Unfortunately, if the d_i are distinct, the only invariant regions for (1.2) are products of intervals (see Smoller [20]). Consequently, the spread of solutions of (1.2) outside more general invariant regions for (1.1) leads us to believe that global existence for (1.1) does not imply global existence for (1.2). Although this work does not resolve this problem, conditions similar to those given in Proposition 1.1, and guaranteeing global existence for (1.2), are stated in § 2. The principal portion of these conditions can be stated as follows.

Suppose that $\mathbf{I}$ is an invariant region (possibly unbounded) for (1.2), $v(x,t) = (v_i(x,t))$ solves (1.2) on cl $(\Omega) \times [0, T_{\max})$, and there exists a function $H \in C^2(\mathbf{I}, [0,\infty))$ such that:

(1.3) $\partial^2 H(z)$ is nonnegative definite for all $z \in \mathbf{I}$;

(1.4) $H(z) \to \infty$ as $|z| \to \infty$;

(1.5) There exists $M \in \mathbf{R}$ such that $\nabla H(z) \cdot f(z) \leqq MH(z)$ for all $z \in \mathbf{I}$.

Note that (1.3) and (1.4) assert that H is a convex, coercive function from $\mathbf{I}$ to $[0,\infty)$, and that (1.5) imposes a growth restriction on the vector field f across level sets of H. The additional conditions given in § 2 place a polynomial growth restriction on f, along with an "intermediate sum" restriction that allows us to handle distinct diffusion coefficients d_i. When the diffusion coefficients are not distinct, we have the following simple result.

PROPOSITION 1.2. *If $d_i = d_j$ for all $1 \leqq i, j \leqq m$ and there exists a function $H \in C^2(\mathbf{I}, [0,\infty))$ satisfying (1.3)-(1.5), then $T_{\max} = \infty$. Furthermore, if $M \leqq 0$ then there exists $N > 0$ such that $|v(\cdot,t)|_{\infty,\Omega} \leqq N$ for all $t \geqq 0$.*

Proof. Set $u(x,t) = H(v(x,t))$ on $\Omega \times [0, T_{\max})$. Then from (1.3)-(1.5), $u \geqq 0$ and

$$
\begin{aligned}
&u_t(x,t) \leqq d_1 \Delta u(x,t) + Mu(x,t), && (x,t) \in \Omega \times (0, T_{\max}),\\
&\partial u(x,t)/\partial\eta = 0, && (x,t) \in \partial\Omega \times (0, T_{\max}),\\
&u(x,0) = H(v(x,0)), && x \in \Omega.
\end{aligned} \tag{1.6}
$$

Thus, if $K = |H(v(x, 0))|_{\infty,\Omega}$, then application of the strong maximum principle (see Sperb [21, § 2.3]) implies that $0 \leqq u(x, t) \leqq Ke^{Mt}$ on $\Omega \times [0, T_{\max})$. Hence, from the basic existence theorem given above and (1.4), we have $T_{\max} = \infty$. Furthermore, if $M \leqq 0$, then there exists $N > 0$ such that $|v(\cdot, t)|_{\infty,\Omega} \leqq N$ for all $t \geqq 0$.

We remark that if, in addition to (1.3)-(1.5), H also satisfies $H(z) = 0$ if and only if $z = 0$, then $M < 0$ implies $|v(\cdot, t)|_{\infty,\Omega} \to 0$ as $t \to \infty$.

Conditions (1.3)-(1.5) are actually a generalization (see § 5) of a "dissipativity condition" considered by Groger [8] requiring a function related to the dissipation rate of the chemical reactions to be nonnegative. From the work of Horn, Feinberg, and Jackson [5], [11], [12] on mass-action kinetics in reaction networks, it follows that this condition is satisfied for many systems of practical interest. Also, many standard "energy inequalities" arising in mechanical problems correspond to the choices $H(v) = c_1 v_1^2 + \cdots + c_m v_m^2$ for some positive constants $c_1, \cdots, c_m$. For chemical systems we usually choose quite different H's, as illustrated in § 5.

The material in this paper is organized as follows. The notation and main results are given in § 2. In § 3 we develop some useful a priori bounds, and in § 4 we prove the results stated in § 2. Section 5 contains applications of this theory to some reaction-diffusion and nerve conduction problems.

2. Notation and statements of main results. We assume that the reader is familiar with the standard L^p and Sobolev spaces. If $0 \leqq \tau < T$ and $1 \leqq p < \infty$, then $W^{2,1,p}(\Omega \times (\tau, T))$ will denote the Banach space consisting of the elements u of $L^p(\Omega \times (\tau, T))$, having the distributional derivatives $D_t^r D_x^s u$, where $2r + s \leqq 2$ and each of the derivatives lies in $L^p(\Omega \times (\tau, T))$. The norm is defined by

$$|u|^{(2)}_{p,\Omega\times(\tau,T)} = \sum_{2r+s\leqq 2} |\partial_t^r \partial_x^s u|_{p,\Omega\times(\tau,T)}.$$

If $\alpha, \beta \in \mathbf{R}^m$, then $\alpha \leqq \beta$ if and only if $\alpha_i \leqq \beta_i$ for all $1 \leqq i \leqq m$. Also, the positive orthant of $\mathbf{R}^m$ is defined by $\mathbf{P}^m = \{x \in \mathbf{R}^m : x_i > 0 \text{ for all } 1 \leqq i \leqq m\}$, and the nonnegative orthant of $\mathbf{R}^m$ is defined by $\mathrm{cl}(\mathbf{P}^m)$.

Throughout, Ω will be given as in § 1. Furthermore, if s and t are real numbers satisfying $0 \leqq s < t$, then $\Omega \times [s, t]$ will be denoted by $Q[s, t]$. Similarly, $\Omega \times [s, t)$, $\Omega \times (s, t)$, and $\Omega \times (s, t]$ will be denoted by $Q[s, t)$, $Q(s, t)$, and $Q(s, t]$, respectively.

The primary concern of this work is the system

$$\begin{aligned} &v_t(x, t) = D\Delta v(x, t) + f(v(x, t)), && x \in \Omega, \quad t > 0, \\ &Bv(x, t) = \gamma, && x \in \partial\Omega, \quad t > 0, \\ &v(x, 0) = v_0(x), && x \in \Omega, \end{aligned} \tag{2.1}$$

where D and f are given as in § 1, and B and γ satisfy the following assumptions:

(A1) $\alpha = (\alpha_i)$, $\gamma = (\gamma_i) \in \mathbf{R}^m$ and $B = (B_i)$ is a diagonal operator given by $Bv = (B_i v_i)$, where $B_i v_i = \alpha_i v_i + \beta(\partial v_i / \partial \eta)$ for all $1 \leqq i \leqq m$. Furthermore, α, β, and γ satisfy: (i) $\alpha_k \geqq 0$ and $\beta \in \{0, 1\}$ for all $1 \leqq k \leqq m$; (ii) if $\beta = 0$ then $\alpha_k = 1$ for all $1 \leqq k \leqq m$; and (iii) if $\alpha_i = 0$ for some $1 \leqq i \leqq m$, then $\alpha \equiv 0$, $\beta = 1$, and $\gamma \equiv 0$.

(A2) $v_0 = (v_{i0}) \in L^\infty(\Omega, \mathbf{R}^m)$.

Conditions (A1), (A2) guarantee local existence and uniqueness for (2.1). A proof of the following theorem can be found in Hollis, Martin, and Pierre [10, Prop. 1, p. 745].

THEOREM 2.1. *Suppose that* (A1), (A2) *hold. Then there exists* $T_{\max} > 0$ *and* $N = (N_i) \in C([0, T_{\max}), \mathbf{R}^m)$ *such that*

(i) (2.1) *has a unique, classical, noncontinuable solution* $v(x, t)$ *on* $\mathrm{cl}(\Omega)\times[0, T_{\max})$; *and*

(ii) $|v_i(\cdot, t)|_{\infty,\Omega} \leqq N_i(t)$ *for all* $1 \leqq i \leqq m$, $0 \leqq t < T_{\max}$.

Moreover, if $T_{\max} < \infty$, *then* $|v_i(\cdot, t)|_{\infty,\Omega} \to \infty$ *for some* $1 \leqq i \leqq m$.

For the remainder of this paper $\mathbf{I}$ will be a (possibly unbounded) subset of $\mathbf{R}^m$ for which (2.1) is invariant. Since our primary interest lies in the case when the "diffusion coefficients" d_i are distinct, it follows from our comment in § 1 that there exist (possibly unbounded) intervals $\mathbf{I}_i$ of $\mathbf{R}$ such that $\mathbf{I} = \mathbf{I}_1 \times \cdots \times \mathbf{I}_m$. In addition, this invariance assumption seems to warrant the following.

(A3) If $1 \leqq i \leqq m$ and $\alpha_i \neq 0$, then $\gamma_i/\alpha_i \in \mathbf{I}_i$. Furthermore, $v_0(x) \in \mathbf{I}$ for all x.

In order to state a version of Proposition 1.1 for (2.1), we give a restricted version of (1.3)-(1.5). Suppose $v = (v_i)$ solves (2.1) and there exist functions $H \in C^2(\mathbf{I}, \mathbf{R})$ and $h_i \in C^2(\mathbf{I}_i, \mathbf{R})$ for each $1 \leqq i \leqq m$ such that:

(H1) $H(z) = \sum_{i=1}^m h_i(z_i)$ for all $z \in \mathbf{I}$.

(H2) $h_i(z_i)$, $h_i''(z_i) \geqq 0$ for all $z_i \in \mathbf{I}_i$, $1 \leqq i \leqq m$.

(H3) $H(z) \to \infty$ if and only if $|z| \to \infty$ in $\mathbf{I}$.

(H4) There exists $A = (a_{ij}) \in \mathbf{R}^{m\times m}$ satisfying $a_{ij} \geqq 0$, $a_{ii} > 0$ for all $1 \leqq i, j \leqq m$, such that for each $1 \leqq j \leqq m$, either (i) there exist r, K_1, $K_2 \geqq 0$, independent of j, such that $\sum_{i=1}^j a_{ji} h_i'(z_i) f_i(z) \leqq K_1(H(z))^r + K_2$ for all $z \in \mathbf{I}$; or (ii) there exists $\zeta \geqq 1$ such that for all $\zeta \leqq p < \infty$ there exist $0 < \delta_p < 1$ and K_{3p}, $K_{4p} \in C([0, \infty))$ such that for all $0 \leqq \tau < T < T_{\max}$ we have $|h_j(v_j)|_{p,Q[\tau,T]} \leqq K_{3p}(T-\tau) + K_{4p}(T-\tau)|H(v)|^{\delta_p}_{p,Q[\tau,T]}$.

(H5) There exist q_1, K_5, $K_6 \geqq 0$ such that for all $1 \leqq i \leqq m$ we have $h_i'(z_i) f_i(z) \leqq K_5(H(z))^{q_1} + K_6$ for all $z \in \mathbf{I}$.

(H6) There exist K_7, $K_8 \geqq 0$ such that $\nabla H(z) \cdot f(z) \leqq K_7 H(z) + K_8$ for all $z \in \mathbf{I}$.

Remarks. (i) Conditions (H1)-(H3) and (H6) are a restatement of (1.3)-(1.5) with a "splitting condition" imposed on H. We show in § 3 that these conditions are sufficient to obtain certain a priori bounds on $H(v)$.

(ii) Hypothesis (H4) requires that either there exists a polynomial upper bound for the jth "intermediate sum," or $h_j(v_j)$ satisfies an L^p growth restriction. Note that (H4)(ii) is satisfied if v_j can be bounded a priori on $Q(0, T_{\max})$. The results we obtain are dependent on r if (H4)(ii) does not always hold, but the restrictions on r allow nontrivial nonlinearities in f. We show in § 4 that if (H4)(ii) holds for all $1 \leqq j \leqq m$, then $T_{\max} = \infty$.

(iii) Hypothesis (H5) is essentially a polynomial growth restriction on f. The results we obtain are independent of the size of q_1.

In § 5, we will see that several model systems satisfy (H1)-(H6). Our first result is stated as Theorem 2.2.

THEOREM 2.2. *Suppose* (A1)-(A3) *and* (H1)-(H5) *hold. If there exist* $a > 0$ *and* $g \in C([0, \infty))$ *such that*

$$\left| \int_0^t (H(v(\cdot, s)))^a \, ds \right|_{\infty,\Omega} \leqq g(t) \quad \text{for all } 0 < t < T_{\max}, \tag{2.2}$$

and $r < 1 + a$, *then* $T_{\max} = \infty$.

Although it is not obvious, Theorem 2.2 gives us a generalization of Proposition 1.2 to cases of general diffusion coefficients and more general boundary conditions. Suppose that (A1)-(A3), (H1), (H2), and (H6) are satisfied. Then it will be shown in § 3 that (2.2) holds with $a = 1$, independent of m and n! Hence, if $r < 2$ in (H4)(i), then this theorem provides a global existence result independent of space dimension and size of diffusion coefficients. Thus, returning to consider systems (1.1) and (1.2),

we see that global existence for (1.1), with an appropriate H, implies global existence for (1.2) for a large class of nontrivial systems.

Theorem 2.2 also generalizes the global existence result of Hollis, Martin, and Pierre, who considered systems of the form (2.1) with $m=2$ and the following assumptions:

(a) $f_1(0, z_2), f_2(z_1, 0) \geqq 0$ for all $z_1, z_2 \geqq 0$.

(b) $v_{i0}(x) \geqq 0$, for all $x \in \Omega$, $i=1,2$.

(c) $f_1(v(x,t))+f_2(v(x,t)) \leqq M(v_1(x,t))$, for some $M \in C([0,\infty))$ on $Q(0, T_{\max})$.

(d) There exists $g \in C([0,\infty))$ such that $|v_1(\cdot, t)|_{\infty,\Omega} \leqq g(t)$ for all $0<t<T_{\max}$.

Assumptions (a) and (b) above imply that cl $(\mathbf{P}^2)$ is invariant for (2.1) (see Lightbourne and Martin [15]), whereas (c) and (d) are simple cases of (H6) and (H4)(ii), respectively, with $H(v)=v_1+v_2+$constant. Hence, these assumptions imply that the hypotheses of Theorem 2.2 hold.

We can also extend a version of a very general result of Amann's [2] (which was given for nonlinear parabolic systems) as it applies to systems of the form (2.1). Amann's result (as it applies to (2.1)) can be stated as follows. Suppose that:

(i) There exist K, $r>0$ such that for all $1 \leqq j \leqq m$

$$|f_j(v(x,t))| \leqq K\left(\sum_{i=1}^{m} v_i(x,t)+1\right)^r \quad \text{on } Q(0, T_{\max});$$

(ii) There exist $g \in C([0,\infty))$ and $a>0$ *such that for all* $1 \leqq i \leqq m$

$$\left|\int_\Omega |v_i(x,\cdot)|^a \, dx\right|_{\infty,(0,t)} \leqq g(t) \quad \text{for all } 0<t<T_{\max};$$

(iii) $r<1+2a/n$;

then $T_{\max}=\infty$. We state our extension as Theorem 2.3.

THEOREM 2.3. *Suppose* (A1)-(A3) *and* (H1)-(H5) *hold. If there exist* $a>0$ *and* $g \in C([0,\infty))$ *such that*

$$\left|\int_\Omega (H(v(x,\cdot)))^a \, dx\right|_{\infty,(0,t)} \leqq g(t) \quad \textit{for all } 0<t<T_{\max},$$

$$r<\begin{cases} 1+2a/n, & n \geqq 2, \\ 1+a/n, & n=1, \end{cases} \tag{2.3}$$

then $T_{\max}=\infty$.

Note that assumption (H4)(i) allows the possibility of higher-order terms cancelling in "intermediate-sums." Hence, an $L^a(\Omega)$ a priori bound for the solution of (2.1) might be used more effectively in Theorem 2.3 than in Amann's result. Also, if (A1)-(A3), (H1), (H2), and (H6) hold, then (as shown in § 3) (2.3) holds with $a=1$. Consequently, Theorem 2.3 can be applied to many nontrivial systems.

Global existence results for (2.1) can also be given in terms of $L^a(Q(0,T))$ a priori bounds.

THEOREM 2.4. *Suppose* (A1)-(A3) *and* (H1)-(H5) *are satisfied. If there exist* $a>0$ *and* $g \in C([0,\infty))$ *such that*

$$\int_0^t \int_\Omega (H(v(x,s)))^a \, dx \, ds \leqq g(t) \quad \textit{for all } 0<t<T_{\max} \tag{2.4}$$

and $r<1+2a/(n+2)$, *then* $T_{\max}=\infty$.

As above, we show in § 3 that if (A1)-(A3), (H1), (H2), and (H6) hold, then (2.4) is satisfied with $a=2$. Thus, when $n=1$, quadratic upper bounds on "intermediate sums" are permitted.

3. A priori estimates. In this section, we show that solutions of (2.1) satisfy certain norm bounds if conditions (A1)-(A3), (H1)-(H3), and (H6) are satisfied. It should be noted that for a particular system, better a priori bounds might be obtainable than the ones given here, but these will at least indicate the merit of Theorems 2.2-2.4.

The following technical lemma will be useful in this section and § 5. We omit the straightforward proof.

LEMMA 3.1. *Suppose that* (A1), (A3), *and* (H1)-(H3) *are satisfied. Then there exist* $\delta=(\delta_i)$, $\sigma=(\sigma_i)\in R^m$, *and* $\xi\in\{0,1\}$ *such that* $\delta_i h_i(v_i(x,t))+\xi\,\partial(h_i(v_i(x,t)))/\partial\eta\leqq\sigma_i$ *on* $\partial\Omega\times(0,T_{\max})$ *for all* $1\leqq i\leqq m$. *Furthermore, parts* (i)-(iii) *of* (A1) *are satisfied with* δ, σ, *and* ξ *in place of* α, β, *and* γ, *respectively.*

Throughout the remainder of this section, $h_i(v_i(x,t))$, $h_i'(v_i(x,t))f_i(v(x,t))$, and $h_i(v_{i0}(x))$ will be denoted by $u_i(x,t)$, $F_i(x,t)$, and $u_{i0}(x)$, respectively.

The first result in this section yields bounds for $H(v)$ in $L^1(0,T)$.

THEOREM 3.2. *Suppose that* (A1)-(A3), (H1), (H2), *and* (H6) *are satisfied. Then there exists* $g\in C([0,\infty))$ *such that*

$$\|H(v(\cdot,\cdot))|_{1,(0,t)}|_{\infty,\Omega}\leqq g(t)\quad \textit{for all } 0<t<T_{\max}.$$

Proof. From (2.1), (H1), (H2), and (H6), we have

$$\nabla H(v(x,t))\cdot v_t(x,t)\leqq\nabla H(v(x,t))\cdot D\Delta v(x,t)+K_7H(v(x,t))+K_8$$

$$\leqq\sum_{i=1}^{m}d_i\Delta u_i(x,t)+K_7H(v(x,t))+K_8\quad\text{on } Q(0,T_{\max}).$$

Thus, if $0<\tau<T_{\max}$, then integration with respect to time yields

$$H(v(x,t))\leqq\Delta\int_\tau^t\sum_{i=1}^{m}d_iu_i(x,s)\,ds+H(v(x,\tau))+K_7\int_\tau^t H(v(x,s))\,ds \tag{3.1}$$

$$+K_8(t-\tau)\qquad\text{on } Q(\tau,T_{\max}).$$

Set $d=\max\{d_1,\cdots,d_m\}$, $M=K_7d/\min\{d_1,\cdots,d_m\}$, $L=|H(v(\cdot,\tau))|_{\infty,\Omega}$, and $w(x,t)=\int_\tau^t\sum_{i=1}^m(d_i/d)u_i(x,s)\,ds$ on $Q(\tau,T_{\max})$. Then

$$w_t(x,t)\leqq d\Delta w(x,t)+L+Mw(x,t)+K_8(t-\tau)\quad\text{on } Q(\tau,T_{\max}). \tag{3.2}$$

Referring to Lemma 3.1, set $b=\min\{\delta_1,\cdots,\delta_m\}$, $c=\xi$, and $e=\max\cdot\{\sigma_1,\cdots,\sigma_m\}$. Then

$$bw(x,t)+c\frac{\partial w(x,t)}{\partial\eta}\leqq e(t-\tau)\quad\text{on } \partial\Omega\times(\tau,T_{\max}). \tag{3.3}$$

Furthermore,

$$w(x,\tau)\equiv 0\quad\text{on } \Omega. \tag{3.4}$$

Consequently, application of the strong maximum principle for parabolic equations (see Sperb [21, § 2.3]) to (3.2)-(3.4), along with part (ii) of Theorem 2.1, implies that there exists $g\in C([0,\infty))$ such that $|w(\cdot,t)|_{\infty,\Omega}\leqq g(t)$ for all $0<t<T_{\max}$.

The next result yields $L^1(\Omega)$ a priori bounds for the solution of (2.1).

THEOREM 3.3. *Suppose that* (A1)-(A3), (H1)-(H3), *and* (H6) *are satisfied. If* $\sigma\equiv 0$ *or* $\xi\neq 0$ *in Lemma* 3.1, *then there exists* $g\in C([0,\infty))$ *such that* $|H(v(\cdot,t))|_{1,\Omega}\leqq g(t)$ *for all* $0<t<T_{\max}$.

Proof. Integrating (4.1) over Ω with $\tau = 0$, applying integration by parts, and setting $L_1 = |\Omega||H(v_0)|_{\infty,\Omega}$, we obtain

$$|H(v(\cdot, t))|_{1,\Omega} \leqq \int_0^t \int_{\partial\Omega} \sum_{i=1}^{m} d_i \frac{\partial u_i(x, s)}{\partial \eta}\, d\Sigma\, ds + L_1 + K_7 \int_0^t |H(v(\cdot, s))|_{1,\Omega}\, ds + K_8|\Omega| t \tag{3.5}$$

for all $0 < t < T_{\max}$, with $d\Sigma$ denoting the usual surface measure on $\partial\Omega$. Thus, if $\sigma \equiv 0$ or $\xi \neq 0$, then $\partial u_i(x, s)/\partial \eta \leqq \sigma_i$ for all $1 \leqq i \leqq m$ on $\partial\Omega \times (0, T_{\max})$. Hence, if we set $L_2 = \sum_{i=1}^{m} (d_i \sigma_i |\partial\Omega| + K_8|\Omega|)$, we obtain

$$|H(v(\cdot, t))|_{1,\Omega} \leqq L_2 t + L_1 + K_7 \int_0^t |H(v(\cdot, s))|_{1,\Omega}\, ds \quad \text{for all } 0 < t < T_{\max}. \tag{3.6}$$

Application of Gronwall's inequality to (3.6) proves the result.

As is mentioned in § 2, $L^2(Q(0, T))$ norm bounds are also obtainable for solutions of (2.1). We state this result as Theorem 3.4.

THEOREM 3.4. *Suppose that* (A1)-(A3), (H1)-(H3), *and* (H6) *are satisfied. If* $\sigma \equiv 0$ *or* $\xi \neq 0$ *in Lemma* 3.1, *then there exists* $g \in C([0, \infty))$ *such that* $|H(v)|_{2,Q(0,t)} \leqq g(t)$ *for all* $0 < t < T_{\max}$.

Proof. Choose $d > \max\{d_1, \cdots, d_m\}$ and $0 < \tau < T_{\max}$. For $1 \leqq i \leqq m+1$, let z_i be the solution of

$$\begin{aligned} &z_{it}(x, t) = d_i \Delta z_i(x, t) + F_i(x, t) \quad \text{on } Q(\tau, T_{\max}), \\ &\delta_i z_i(x, t) + \xi \frac{\partial z_i(x, t)}{\partial \eta} = \sigma_i \quad \text{on } \partial\Omega \times (\tau, T_{\max}), \\ &z_i(x, \tau) = L_i \quad \text{on } \Omega \end{aligned} \tag{3.7}$$

with $F_{m+1}(x, t) = K_7 H(v(x, t)) + K_8 - \nabla H(v(x, t)) \cdot f(v(x, t))$, $d_{m+1} = d_m$, $\delta_{m+1} = \delta_m$, $\sigma_{m+1} = \sigma_m$, and $L_{m+1} = L_m$, where $L_i = |u_i(\cdot, \tau)|_{\infty,\Omega}$ for all $1 \leqq i \leqq m$. Then, from the strong maximum principle for parabolic equations, we have $z_i(x, t) \geqq u_i(x, t)$ for all $1 \leqq i \leqq m$, and there exists $g_1 \in C([0, T_{\max}))$ such that $0 \leqq |z_i(\cdot, t)|_{\infty,\Omega} \leqq g_1(t)$ for all $\tau \leqq t < T_{\max}$.

Proceeding as in the proof of Theorem 3.2, set

$$w(x, t) = \int_\tau^t \sum_{i=1}^{m} \frac{d_i}{d} z_i(x, s)\, ds \quad \text{on } Q[\tau, T_{\max}),$$

$$b = \min\{\delta_1, \cdots, \delta_m\}, \quad c = \xi, \quad e = \max\{\sigma_1, \cdots, \sigma_m\}.$$

Then

$$\begin{aligned} w_t(x, t) &= d\Delta w(x, t) + \sum_{i=1}^{m+1} z_i(x, \tau) + \sum_{i=1}^{m+1} \left(\frac{d_i}{d} - 1\right) z_i(x, t) + K_7 \int_\tau^t H(v(x, s))\, ds \\ &\quad + K_8(t - \tau) \qquad \text{on } Q(\tau, T_{\max}), \\ bw(x, t) + c\frac{\partial w(x, t)}{\partial \eta} &\leqq e(t - \tau) \quad \text{on } \partial\Omega \times (\tau, T_{\max}), \\ w(x, \tau) &= 0 \quad \text{on } \Omega. \end{aligned}$$

Note that there exists $g_2 \in C([0, \infty))$ such that $|w(\cdot, t)|_{\infty,\Omega} \leqq g_2(t)$ for all $\tau < t < T_{\max}$, where g_2 satisfies the same properties as the function g given in Theorem 3.2. Similarly,

for each $1 \leqq i \leqq m$, z_i satisfies the same $L^1(\Omega)$ norm bounds as $H(v)$ in Theorem 3.3. Also, from above, for $\tau < t < T_{\max}$ and $1 \leqq i \leqq m+1$, we have

$$
\begin{aligned}
&\int_\tau^t \int_\Omega z_i \sum_{j=1}^{m+1} \left(\frac{d_j}{d} - 1\right) z_j \, dx \, ds \\
&\quad = \int_\tau^t \int_\Omega z_i \left[w_t - d\Delta w - G - K_7 \int_\tau^s H(v(x,r))\, dr \right] dx \, ds \\
&\quad = \int_\tau^t \int_\Omega \left\{ z_i \left[w_t - G - K_7 \int_\tau^s H(v(x,r))\, dr \right] - dw\, \Delta z_i \right\} dx \, ds \\
&\qquad - d \int_\tau^t \int_{\partial\Omega} \left(z_i \frac{\partial w}{\partial \eta} - w \frac{\partial z_i}{\partial \eta} \right) d\Sigma \, ds,
\end{aligned} \tag{3.8}
$$

where $G(x,t) = \sum_{i=1}^{m+1} z_i(x,\tau) + K_8(t-\tau)$ on $Q(\tau, T_{\max})$.

Now,

$$
\begin{aligned}
\int_\tau^t \int_\Omega dw\, \Delta z_i \, dx \, ds &= \int_\tau^t \int_\Omega \frac{d}{d_i} w(z_{it} - F_i)\, dx \, ds \\
&= \int_\Omega \frac{d}{d_i} w(x,t) z_i(x,t)\, dx - \int_\tau^t \int_\Omega \frac{d}{d_i} (z_i w_t + w F_i)\, dx \, ds.
\end{aligned}
$$

Making this substitution in (3.8), noting that $w_t(x,t) = \sum_{j=1}^{m+1} (d_j/d) z_j(x,t)$, and combining terms, we obtain

$$
\begin{aligned}
&\int_\tau^t \int_\Omega \left(\sum_{i=1}^{m+1} \frac{d_i}{d} z_i \right) \left(\sum_{j=1}^{m+1} \left(\frac{d_j}{d} - 1 \right) z_j \right) dx \, ds \\
&\quad = \int_\tau^t \int_\Omega \left(\sum_{i=1}^{m+1} \left(\frac{d_i}{d} + 1 \right) z_i \right) \left(\sum_{j=1}^{m+1} \frac{d_j}{d} z_j \right) dx \, ds \\
&\qquad - \int_\Omega w(x,t) \sum_{i=1}^{m+1} z_i(x,t)\, dx \\
&\qquad - \int_\tau^t \int_\Omega \sum_{i=1}^{m+1} \frac{d_i}{d} z_i \left[G + K_7 \int_\tau^t H(v(x,r))\, dr \right] dx \, ds \\
&\qquad + \int_\tau^t \int_\Omega w(K_7 H(v) + K_8)\, dx \, ds \\
&\qquad - \int_\tau^t \int_{\partial\Omega} \sum_{i=1}^{m+1} d_i \left[z_i \frac{\partial w}{\partial \eta} - w \frac{\partial z_i}{\partial \eta} \right] d\Sigma \, ds.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
2 \int_\tau^t \int_\Omega \sum_{i=1}^{m+1} \frac{d_i}{d} (z_i)^2 \, dx \, ds &= \int_\Omega w(x,t) \sum_{i=1}^{m+1} z_i(x,t)\, dx - \int_\tau^t \int_\Omega w(K_7 H(v) + K_8)\, dx \, ds \\
&\quad + \int_\tau^t \int_\Omega \sum_{i=1}^{m+1} \frac{d_i}{d} z_i \left[G + K_7 \int_\tau^t H(v(x,r))\, dr \right] dx \, ds \\
&\quad + \int_\tau^t \int_{\partial\Omega} \sum_{i=1}^{m+1} d_i \left[z_i \frac{\partial w}{\partial \eta} - w \frac{\partial z_i}{\partial \eta} \right] d\Sigma \, ds.
\end{aligned} \tag{3.9}
$$

If $\delta \equiv \xi = 0$, then the fourth term on the right-hand side of (3.9) is identically zero. If $\xi \neq 0$, then the fourth term on the right-hand side of (3.9) can be bounded above by $\int_\tau^t \int_{\partial\Omega} d_i z_i(e+\delta_i w)\, d\Sigma\, ds$. Thus, if we apply Hölder's inequality and Theorems 3.2 and 3.3 to the right-hand side of (3.9), then the result follows.

4. Proofs of global existence results. Throughout this section, we assume that (A1)-(A3) and (H1)-(H5) hold. Suppose $v(x, t)$ solves (2.1) on $\mathrm{cl}(\Omega)\times[0, T_{\max})$, $0 \leqq \tau < T < T_{\max}$, and δ, σ, and ξ are given by Lemma 3.1. Furthermore, suppose $1 < p < \infty$, $\theta \in L^p(Q(\tau, T))$ satisfies $|\theta|_{p,Q(\tau,T)} = 1$ and $\theta \geqq 0$, and for $1 \leqq i \leqq m$, Ψ_i is the solution of

$$
\begin{aligned}
&\Psi_{it}(x, t) = -d_i \Delta \Psi_i(x, t) - \theta(x, t) && \text{on } Q[\tau, T),\\
&b_1 \Psi_i(x, t) + \xi \partial \Psi_i(x, t)/\partial\eta = 0 && \text{on } \partial\Omega \times [\tau, T),\\
&\Psi_i(x, T) = 0 && \text{on } \Omega,
\end{aligned} \tag{4.1}
$$

where $b_1 = \min\{\delta_1, \cdots, \delta_m\}$.

Note that if $w_i(x, t) = \Psi_i(x, T-t)$ on $\mathrm{cl}(\Omega)\times[0, T-\tau]$, then

$$
\begin{aligned}
&w_{it}(x, t) = d_i \Delta w_i(x, t) + \theta(x, T-t) && \text{on } Q(0, T-\tau],\\
&b_1 w_i(x, t) + \xi \partial w_i(x, t)/\partial\eta = 0 && \text{on } \partial\Omega \times (0, T-\tau],\\
&w_i(x, 0) = 0 && \text{on } \Omega.
\end{aligned} \tag{4.2}
$$

Thus, the well-known results (cf. Ladyzenskaja, Solonnikov, and Uralceva [13, Thm. 9.1, p. 341]) and basic maximum principles imply that $\Psi_i \in W^{2,1,p}(Q(\tau, T))$ and $\Psi_i \geqq 0$. Furthermore, we have the following important lemma.

LEMMA 4.1. *For all $1 < p < \infty$ and $1 \leqq i, j \leqq m$, there exists $C_{p(T-\tau)} > 0$ (independent of θ) such that*

(i) $|\Psi_i(\tau)|_{p,\Omega}, |\Psi_i|^{(2)}_{p,Q(\tau,T)} \leqq C_{p(T-\tau)}$,

(ii) $$\int_\tau^T \int_{\partial\Omega} \left[\Psi_j(x, t) \frac{\partial(h_i(v_i(x, t)))}{\partial\eta} - h_i(v_i(x, t)) \frac{\partial \Psi_j(x, t)}{\partial\eta} \right] d\Sigma\, ds \leqq C_{p(T-\tau)}.$$

Furthermore, if $1 < p < (n+2)/2$ and $q = p(n+2)/(n+2-2p)$ then there exists $K_{q(T-\tau)}$ such that

(iii) $|\Psi_i|_{q,Q(\tau,T)} \leqq K_{q(T-\tau)}$.

Proof. Parts (i) and (iii) follow immediately from (4.1), (4.2), [13, Thm. 9.1, p. 341], and [10, Lemma 2, p. 750]. For part (ii), note that we have $\xi \in \{0, 1\}$. Let $C_{p(T-\tau)}$ be given to satisfy part (i). If $\xi = 1$, then

$$
\begin{aligned}
\int_\tau^T \int_{\partial\Omega} \left[\Psi_i \frac{\partial(h_i(v_i))}{\partial\eta} - h_i(v_i) \frac{\partial \Psi_j}{\partial\eta} \right] d\Sigma\, ds &\leqq \int_\tau^T \int_{\partial\Omega} (\Psi_j(\sigma_i - \delta_i h_i(v_i)) + b_1 h_i(v_i))\, d\Sigma\, ds\\
&\leqq \int_\tau^T \int_{\partial\Omega} \sigma_i \Psi_j\, d\Sigma\, ds\\
&\leqq \sigma_i |\partial\Omega|^{(p-1)/p} \int_\tau^T |\Psi(\cdot, s)|_{p,\partial\Omega}\, ds\\
&\leqq \sigma_i |\partial\Omega|^{(p-1)/p} \int_\tau^T L_p |\Psi_j(\cdot, s)|_{2,p,\Omega}\, ds\\
&\leqq \sigma_i |\partial\Omega|^{(p-1)/p} L_p C_{p(T-\tau)} (T-\tau)^{(p-1)/p},
\end{aligned}
$$

from the definition of b_1, part (i), Hölder's inequality, and the Trace Class Imbedding Theorem (cf. Grisvard [7, Chap. 1, Thm. 1.5.1.2]).

Similarly, if $\xi = 0$, then $\delta_i = 1$ for all $1 \leqq i \leqq m$. Thus, $\Psi_j \equiv 0$ and $\partial \Psi_j / \partial \eta \leqq 0$ on $\partial\Omega \times (\tau, T)$, for all $1 \leqq j \leqq m$. Therefore,

$$\int_\tau^T \int_{\partial\Omega} \left[\Psi_j \frac{\partial (h_i(v_i))}{\partial \eta} - h_i(v_i) \frac{\partial \Psi_j}{\partial \eta} \right] d\Sigma \, ds \leqq \int_\tau^T \int_{\partial\Omega} \sigma_i \left| \frac{\partial \Psi_j}{\partial \eta} \right| d\Sigma \, ds \leqq M_{p(T-\tau)}$$

as above. The result follows.

The following lemma is fundamental to all of the results in this section.

LEMMA 4.2. *Suppose that for all* $1 \leqq j \leqq m$, (H4)(ii) *holds. Then* $T_{\max} = \infty$.

Proof. Suppose (by way of contradiction) that $T_{\max} < \infty$. Consequently, by hypothesis, for all $1 \leqq p < \infty$ there exist M_p, $N_p > 0$, and $0 < \delta_p < 1$ such that

$$|H(v)|_{p,Q(\tau,T)} \leqq M_p + N_p |H(v)|^{\delta_p}_{p,Q(\tau,T)}. \tag{4.3}$$

Thus, $|H(v)|_{p,Q(\tau,T)} < \infty$ for all $1 \leqq p < \infty$. It then follows that

$$|K_5(H(v))^{q_1} + K_6|_{p,Q(\tau,T)} < \infty \quad \text{for all } 1 \leqq p < \infty, \tag{4.4}$$

where K_5, K_6, and q_1 are given in (H5). Now, let $1 \leqq j \leqq m$, set $w = h_j(v_j)$, and suppose $M = |H(v_0)|_{\infty,\Omega}$. Then from the convexity of h_j, Lemma 3.1, and standard maximum principles we have $w \leqq z$, where z solves

$$\begin{aligned} &z_t = d_j \Delta z + K_5 (H(v))^{q_1} + K_6 \quad \text{on } \Omega \times (0, T_{\max}), \\ &\delta_j z + \xi \frac{\partial z}{\partial \eta} = \sigma_j \quad \text{on } \partial\Omega \times (0, T_{\max}), \\ &z = M \quad \text{on } \Omega \times \{0\}. \end{aligned} \tag{4.5}$$

Furthermore, from [13, Thm. 9.1, p. 341] and the Sobolev Imbedding Theorem we have the existence of $\bar{M} > 0$ such that $|z(\cdot, t)|_{\infty,\Omega} < \bar{M}$ for all $0 \leqq t < T_{\max}$. That is, $|h_i(v_i(\cdot, t))|_{\infty,\Omega} < \bar{M}$ for all $0 \leqq t < T_{\max}$. Hence, from (H3), there exists $N > 0$ such that $|v_j(\cdot, t)|_{\infty,\Omega} < N$ for all $0 \leqq t < T_{\max}$ and $1 \leqq j \leqq m$. This contradicts (via Theorem 2.1) our assumption that $T_{\max} < \infty$, and therefore $T_{\max} = \infty$.

Throughout the remainder of this section, we will denote $h_i(v_i(x, t))$, $h_i'(v_i(x, t)) f_i(v(x, t))$, and $h_i(v_{i0}(x))$ by $u_i(x, t)$, $F_i(x, t)$, and $u_{i0}(x)$, respectively.

Proof of Theorem 2.2. Suppose there exists $1 \leqq j \leqq m$ such that u_j does not satisfy (H4)(ii), and if $1 \leqq i \leqq m$ such that $i < j$, then u_i satisfies (H4)(ii). Then (H4)(i) is satisfied for j.

Let $\tau = 0$ in (4.1). Then for $i \leqq j$,

$$\begin{aligned} \int_0^T \int_\Omega u_i \theta \, dx \, ds &= \int_0^T \int_\Omega u_i (-\Psi_{jt} - d_j \Delta \Psi_j) \, dx \, ds \\ &\leqq -\int_0^T \int_\Omega u_i \Psi_{jt} \, dx \, ds - \frac{d_j}{d_i} \int_0^T \int_\Omega \Psi_j d_i \Delta u_i \, dx \, ds \\ &\quad + C_{pT} \quad \text{from Lemma 4.1.} \end{aligned} \tag{4.6}$$

Now, from (H2) and (2.1),

$$\begin{aligned} -\int_0^T \int_\Omega \Psi_j \, d_i \Delta u_i \, dx \, ds &\leqq \int_0^T \int_\Omega \Psi_j (F_i - u_{it}) \, dx \, ds \\ &= \int_0^T \int_\Omega (\Psi_j F_i + u_i \Psi_{jt}) \, dx \, ds + \int_\Omega \Psi_j(x, 0) u_{i0}(x) \, dx. \end{aligned} \tag{4.7}$$

Making this substitution in (4.6) yields

$$\int_0^T \int_\Omega u_i\theta \, dx \, ds \leqq \left(\frac{d_j}{d_i}-1\right) \int_0^T \int_\Omega u_i \Psi_{jt} \, dx \, ds + \frac{d_j}{d_i}\left(\int_0^T \int_\Omega \Psi_j F_i \, dx \, ds + \int_\Omega \Psi(x,0) u_{i0}(x) \, dx\right) + C_{pT}. \tag{4.8}$$

Consequently, if we apply (H4)(i), then

$$\int_0^T \int_\Omega \sum_{i=1}^{j} \frac{d_i}{d_j} a_{ji} u_i \theta \, dx \, ds \leqq \sum_{i=1}^{j-1} a_{ji}\left(1-\frac{d_i}{d_j}\right) \int_0^T \int_\Omega u_i \Psi_{jt} \, dx \, ds + \int_0^T \int_\Omega \Psi_j [K_1 (H(v))^r + K_2] \, dx \, ds + \sum_{i=1}^{j} a_{ji}\left[\int_\Omega \Psi_j(x,0) u_{i0}(x) \, dx + \frac{d_i}{d_j} C_{pT}\right]. \tag{4.9}$$

Now, if we apply Hölder's inequality, Lemma 4.1, and (H4)(ii) for $i<j$ to the first term on the right-hand side of (4.9), then

$$\sum_{i=1}^{j-1} a_{ji}\left(1-\frac{d_i}{d_j}\right) \int_0^T \int_\Omega u_i \Psi_{jt} \, dx \, ds \leqq \sum_{i=1}^{j-1} a_{ji} \left|1-\frac{d_i}{d_j}\right| C_{pT}(K_{3q}(T) + K_{4q}(T)|H(v)|^{\delta q}_{q,Q(0,T)}), \tag{4.10}$$

where $1/p+1/q=1$.

Note that from Lemma 4.1, $|\Psi_{jt}(x,\cdot)|_{p,(0,T)}$ exists for almost all $x \in \Omega$. Thus, if we apply the Sobolev Imbedding Theorem, then there exists $L_{pT}>0$ such that

$$|\Psi_j(x,\cdot)|_{\infty,(0,T)} < L_{pT} |\Psi_j(x,\cdot)|_{1,p,(0,T)} \quad \text{for almost all } x \in \Omega. \tag{4.11}$$

Hence,

$$||\Psi_j(x,\cdot)|_{\infty,(0,T)}|_{p,\Omega} \leqq L_{pT} |\Psi_j|^{(2)}_{p,Q(0,T)} \leqq L_{pT} C_{pT} \quad \text{from Lemma 4.1.} \tag{4.12}$$

So, applying Hölder's inequality and (4.12), we have

$$\begin{aligned} \int_0^T \int_\Omega \Psi_j (H(v))^r \, dx \, ds &= \int_\Omega \int_0^T \Psi_j (H(v))^r \, ds \, dx \\ &\leqq \int_\Omega |\Psi_j(x,\cdot)|_{\infty,(0,T)} \int_0^T (H(v))^r \, ds \, dx \\ &\leqq ||\Psi_j(\cdot,\cdot)|_{\infty,(0,T)}|_{p,\Omega} \left| \int_0^T (H(v(\cdot,s)))^r \, ds \right|_{q,\Omega} \\ &\leqq L_{pT} C_{pT} \left| \int_0^T (H(v(\cdot,s)))^r \, ds \right|_{q,\Omega}. \end{aligned} \tag{4.13}$$

Without loss of generality, assume $r>0$. If $r \leqq a$, then

$$\left| \int_0^T (H(v(\cdot,s)))^r \, ds \right|_{q,\Omega} \leqq |\Omega|^{1/q} T^{(a-r)/a} (g(T))^{r/a}. \tag{4.14}$$

Now suppose $a<r<1+a$. If p is sufficiently close to 1 and $k=a[p-r(p-1)]/[p-a(p-1)]$, then $a/k>1$, $a(r-k)/(a-k)=p/(p-1)$, and $p(a-k)/[a(p-1)]<1$. Thus,

$$\left|\int_0^T (H(v(\cdot,s)))^r\,ds\right|_{q,\Omega} = \left|\int_0^T (H(v(\cdot,s)))^{r-k}(H(v(\cdot,s)))^k\,ds\right|_{q,\Omega} \leqq (g(T))^{k/a}|H(v)|^{p(a-k)/a(p-1)}_{q,Q(0,T)}|\Omega|^{(pk-a)/ap}. \tag{4.15}$$

If we now apply (4.10)-(4.15) and Lemma 4.1 to the right-hand side of (4.9), then

$$\begin{aligned}\int_0^T\int_\Omega \sum_{i=1}^{j} a_{ji}\frac{d_i}{d_j}u_i\theta\,dx\,ds &\leqq \sum_{i=1}^{j-1} a_{ji}\left|1-\frac{d_i}{d_j}\right| C_{pT}(K_{3q}(T)+K_{4q}(T)|H(v)|^{\delta q}_{q,Q(0,T)}) \\ &\quad + K_1L_{pT}C_{pT}[|\Omega|^{1/q}T^{(a-r)/a}g(T) \\ &\qquad + |\Omega|^{(pk-a)/ap}(g(T))^{k/a}|H(v)|^{p(a-k)/a(p-1)}_{q,Q(0,T)}] \\ &\quad + \sum_{i=1}^{j} a_{ji}C_{pT}\left[|u_{i0}|_{q,Q(0,T)}+\frac{d_i}{d_j}\right].\end{aligned} \tag{4.16}$$

Then, since $0<\delta_q$, $p(a-k)/a(p-1)<1$, we see that there exist K_{7q}, $K_{8q}\in C([0,\infty))$ and $0<\varepsilon_q<1$ for all q sufficiently large such that

$$|u_i|_{q,Q(0,T)} \leqq K_{7q}(T)+K_{8q}(T)|H(v)|^{\varepsilon_q}_{q,Q(0,T)} \quad \text{for all } 0<T<T_{\max}. \tag{4.17}$$

It follows that (H4)(ii) holds for j. Therefore we have a contradiction, and hence (H4)(ii) is satisfied for all $1\leqq i\leqq m$. Consequently, the result follows from Lemma 4.2.

Proof of Theorem 2.3. As in the proof of Theorem 2.2, suppose there exists $1\leqq j\leqq m$ such that (H4)(ii) is not satisfied for j, and if $1\leqq i\leqq m$ such that $i<j$, then (H4)(ii) is satisfied for i. Then (H4)(i) is satisfied for j. Setting $\tau=0$ in (4.1), we see that (4.6)-(4.10) are satisfied.

Case 1. Suppose $n\geqq 3$. Note that if $p>1$, then $|\Psi_j(\cdot,s)|_{2,p,\Omega}$ exists for almost all $0<s<T$. Also, if $1<p<n/2$, then the Sobolev Imbedding Theorem implies that there exists $M_p>0$ such that if $q=pn/(n-2p)$, then $|\Psi_j(\cdot,s)|_{q,\Omega}\leqq M_p|\Psi_j(\cdot,s)|_{2,p,\Omega}$ for almost all $0<s<T$. Thus,

$$\left(\int_0^T |\Psi_j(\cdot,s)|^p_{q,\Omega}\,ds\right)^{1/p} \leqq M_p\left(\int_0^T |\Psi_j(\cdot,s)|^p_{2,p,\Omega}\,ds\right)^{1/p} \leqq M_pC_{pT}. \tag{4.18}$$

Consequently,

$$\begin{aligned}&\int_0^T\int_\Omega \Psi_j(H(v))^r\,dx\,ds \\ &\quad \leqq \int_0^T\left(\int_\Omega (\Psi_j)^q\,dx\right)^{1/q}\left(\int_\Omega (H(v))^{rpn/((p-1)n+2p)}\,dx\right)^{(p-1)/p+2/n} ds \\ &\quad \leqq \left(\int_0^T |\Psi_j(\cdot,s)|^p_{p,\Omega}\,ds\right)^{1/p} \\ &\qquad \times\left(\int_0^T\left(\int_\Omega (H(v))^{rpn/((p-1)n+2p)}\,dx\right)^{1+(2p/n(p-1))} ds\right)^{(p-1/p)} \\ &\quad \leqq M_pC_{pT}\left(\int_0^T\left(\int_\Omega (H(v))^{rpn/((p-1)n+2p)}\,dx\right)^{1+(2p/n(p-1))} ds\right)^{(p-1/p)}.\end{aligned} \tag{4.19}$$

Note that for $p>1$, $rpn/((p-1)n+2p)<rn/2$. Hence, if $0<r\leqq 2a/n$, then $rn/2\leqq a$, and there exists $\varepsilon_{pn}>0$ such that

$$\int_0^T\int_\Omega \Psi_j(H(v))^r\,dx\,ds \leqq M_pC_{pT}|\Omega|^{\varepsilon_p}\left(\int_0^T (g(s))^{1+2p/n(p-1)}\,ds\right)^{(p-1)/p}. \tag{4.20}$$

So, suppose $2a/n<r<1+2a/n$. Then for $p>1$ and sufficiently close to 1, if $k=ap[n(p-1)(1-r)+2p]/[p(1-a)+a][(p-1)n+2p]$, then $0<k<a$, $(rpn/((p-1)n+2p)-k)(a/(a-k))=(p/p-1)$, and $(1+2p/(n(p-1)))(a-k)/a<1$. Hence,

$$\begin{aligned}
&\left(\int_0^T\left(\int_\Omega (H(v))^{rpn/((p-1)n+2p)}\,dx\right)^{1+2p/(n(p-1))}ds\right)^{(p-1)/p}\\
&\quad=\left(\int_0^T\left(\int_\Omega (H(v))^{rpn/((p-1)n+2p)-k}(H(v))^k\,dx\right)^{1+2p/(n(p-1))}ds\right)^{(p-1)/p}.\\
&\quad\leqq\left(\int_0^T\left(\int_\Omega (H(v))^{p/(p-1)}\,dx\right)^{(1+2p/(n(p-1)))(a-k)/a}\right.\\
&\qquad\left.\times\left(\int_\Omega (H(v))^a\,dx\right)^{(1+2p/(n(p-1)))k/a}ds\right)^{(p-1)/p}\\
&\quad\leqq |g|_{\infty,(0,T)}^{(1+2p/(n(p-1)))(k(p-1)/(ap))}T^{(kn(p-1)+2p(k-a))/(anp)}|H(v)|_{p/(p-1),Q(0,T)}^{(1+2p/(n(p-1)))(a-k)/a}.
\end{aligned}\tag{4.21}$$

Then, as in the proof of Theorem 2.2, (4.10), (4.19), (4.20), and (4.21) combined with (4.9) imply that (H4)(ii) holds for j. Thus, we have a contradiction, and (H4)(ii) holds for all $1\leqq i\leqq m$. Consequently, the result follows from Lemma 4.2.

Case 2. Suppose $n\leqq 2$. As in Case 1, if $p>1$, then $|\Psi_j(\cdot,s)|_{2,p,\Omega}$ exists for almost all $0<s<T$. Then (from the Sobolev Imbedding Theorem) there exists $M_p>0$ such that $|\Psi_j(\cdot,s)|_{\infty,\Omega}\leqq M_p|\Psi_j(\cdot,s)|_{2,p,\Omega}$ for almost all $0<s<T$. Consequently,

$$\left(\int_0^T |\Psi_j(\cdot,s)|_{\infty,\Omega}^p\,ds\right)^{1/p}\leqq M_p\left(\int_0^T |\Psi_j(\cdot,s)|_{2,p,\Omega}^p\,ds\right)^{1/p}\leqq M_pC_{pT}. \tag{4.22}$$

Then, as in Case 1, we have

$$\begin{aligned}
&\int_0^T\int_\Omega \Psi_j(H(v))^r\,dx\,ds\\
&\quad\leqq\int_0^T |\Psi_j(\cdot,s)|_{\infty,\Omega}\int_\Omega (H(v))^r\,dx\,ds\\
&\quad\leqq\left(\int_0^T |\Psi_j(\cdot,s)|_{\infty,\Omega}^p\,ds\right)^{1/p}\left(\int_0^T\left(\int_\Omega (H(v))^r\,dx\right)^{p/(p-1)}ds\right)^{(p-1)/p}\\
&\quad\leqq M_pC_{pT}\left(\int_0^T\left(\int_\Omega (H(v))^r\,dx\right)^{p/(p-1)}ds\right)^{(p-1)/p}.
\end{aligned}\tag{4.23}$$

Now, if $r\leqq a$, then there exists $\delta_r\geqq 0$ such that

$$\left(\int_0^T\left(\int_\Omega (H(v))^r\,dx\right)^{p/(p-1)}ds\right)^{(p-1)/p}\leqq|\Omega|^{\delta_r}\left(\int_0^T (g(s))^{(p-1)r/pa}\,ds\right)^{(p-1)/p}. \tag{4.24}$$

If $a<r<1+a$ and $p>1$ is sufficiently close to 1, then setting $k=a[r(p-1)-p]/(a(p-1)-p)$ implies $a/k>1$, $(a-k)p/(a(p-1))<1$, and $a(r-k)/$

$(a-k)=p/(p-1)$. Hence,

$$\begin{aligned}
&\left(\int_0^T\left(\int_\Omega (H(v))^r\,dx\right)^{p/(p-1)}ds\right)^{(p-1)/p}\\
&\quad=\left(\int_0^T\left(\int_\Omega (H(v))^{r-k}(H(v))^k\,dx\right)^{p/(p-1)}ds\right)^{(p-1)/p}\\
&\quad\leqq\left(\int_0^T\left(\int_\Omega (H(v))^{p/(p-1)}\,dx\right)^{(a-k)p/(a(p-1))}\left(\int_\Omega (H(v))^a\,dx\right)^{kp/(a(p-1))}ds\right)^{(p-1)/p}\\
&\quad\leqq |g|_{\infty,(0,T)}^{k/a}T^{(kp-a)/ap}|H(v)|_{p/(p-1),Q(0,T)}^{(a-k)p/(a(p-1))}.
\end{aligned}\tag{4.25}$$

We can now combine (4.9), (4.10), and (4.23)-(4.25) to show that (H4)(ii) holds for j. Thus, we have a contradiction, and therefore (H4)(ii) holds for all $1\leqq i\leqq m$. Thus, from Lemma 4.2, the result follows.

Proof of Theorem 2.4. As in the proof of Theorem 2.2, suppose there exists $1\leqq j\leqq m$ such that u_j does not satisfy (H4)(ii), and if $1\leqq i\leqq m$ such that $i<j$, then u_i satisfies (H4)(ii). Then (H4)(i) is satisfied for j. Setting $\tau=0$ in (4.1), we see that (4.5)-(4.10) are satisfied. Also, note that from Lemma 4.1, if

$$1<p<\frac{n+2}{2},\qquad q=\frac{p(n+2)}{n+2-2p},$$

then

$$|\Psi_j|_{q,Q(0,T)}\leqq K_{qT}.\tag{4.26}$$

Thus,

$$\begin{aligned}
&\int_0^T\int_\Omega \Psi_j(H(v))^r\,dx\,ds\\
&\quad\leqq |\Psi_j|_{q,Q(0,T)}\left(\int_0^T\int_\Omega (H(v))^{rp(n+2)/((p-1)(n+2)+2p)}\,dx\,ds\right)^{(p-1)/p+2/(n+2)}\\
&\quad\leqq K_{qT}\left(\int_0^T\int_\Omega (H(v))^{rp(n+2)/((p-1)(n+2)+2p)}\,dx\,ds\right)^{(p-1)/p+2/(n+2)}.
\end{aligned}\tag{4.27}$$

Clearly, $p(n+2)/((p-1)(n+2)+2p)<(n+2)/2$. So, if $r\leqq 2a/(n+2)$, then $rp(n+2)/((p-1)(n+2)+2p)<a$. Consequently,

$$\begin{aligned}
&\left(\int_0^T\int_\Omega (H(v))^{rp(n+2)/((p-1)(n+2)+2p)}\,dx\,ds\right)^{(p-1)/p+2/(n+2)}\\
&\qquad\leqq T^{1-rp(n+2)/(a[(p-1)(n+2)+2p])}|H(v)|_{a,Q(0,T)}^{r/a}.
\end{aligned}\tag{4.28}$$

Now, suppose $2a/(n+2)<r<1+2a/(n+2)$. If p is sufficiently close to 1 and we set

$$k=\frac{ap[(p-1)(n+2)+2p]-arp(p-1)(n+2)}{[p-a(p-1)][(p-1)(n+2)+2p]},$$

then $a/k>1$, $(rp(n+2)/((p-1)(n+2)+2p)-k)a/(a-k)=p/(p-1)$, and $((p-1)/p$

$+2/(n+2))(a-k)/a<(p-1)/p$. Hence, there exists $0<\delta_q<1$ such that

$$\left(\int_0^T\int_\Omega (H(v))^{rp(n+2)/((p-1)(n+2)+2p)}\,dx\,ds\right)^{(p-1)/p+2/(n+2)}$$

$$=\left(\int_0^T\int_\Omega (H(v))^{rp(n+2)/((p-1)(n+2)+2p)-k}(H(v))^k\,dx\,ds\right)^{(p-1)/p+2/(n+2)} \tag{4.29}$$

$$\leqq (g(T))^{(k/a)((p-1)/p+2/(n+2))}|H(v)|^{\delta_q}_{q,Q(0,T)}.$$

Then, as in the proof of Theorem 2.2, combining (4.9), (4.10), and (4.27)-(4.29), we see that (H4)(ii) holds for j. This is a contradiction, and therefore (H4)(ii) holds for all $1\leqq i\leqq m$. Therefore, from Lemma 4.2, the result follows.

5. Applications. In this section we show that if (2.1) is a system modeling reaction-diffusion, then conditions (H1)-(H3) and (H6) generalize a "dissipativity condition" used by Groger [8] that requires a function closely related to the dissipation rate of the chemical reactions to be nonnegative. We also use Theorem 2.2 to analyze specific systems modeling certain reaction-diffusion and nerve conduction problems.

Assume that (2.1) is a system modeling reaction-diffusion and $v(x,t)$ solves (2.1). The dissipativity condition considered by Groger can be stated as follows.

There exist $a\in P^m$ and $g\in C(P^m,R)$ such that

$$\lim_{|z|\to\infty} g(z)/|z|=0, \text{ and for all } z\in P^m,\ -\sum_{i=1}^m f_i(z)\ln(z_i/a_i)+g(z)\geqq 0, \tag{5.1}$$

where, as noted by Groger, $-\sum_{i=1}^m f_i(v(x,t))\ln(v_i(x,t)/a_i)$ can be interpreted as a suitably scaled dissipation rate of the chemical reactions.

Now, suppose that P^m is invariant for (2.1), $a\in P^m$, and $H:P^m\to R$ is defined by $H(z)=\sum_{i=1}^m h_i(z_i)$, where $h_i(z_i)=z_i\ln(z_i/a_i)-z_i+a_i$. Then simple calculation shows that h_i, $h_i''\geqq 0$ on P^m, and $H(z)\to\infty$ if and only if $|z|\to\infty$ in P^m. Furthermore, from (5.1) and the definition of H, there exist M, $N>0$ such that $g(z)\leqq M\sum_{i=1}^m z_i+N$ and $H(z)+(\exp(1)-1)\sum_{i=1}^m a_i\geqq\sum_{i=1}^m z_i$ for all $z\in P^m$. Thus, $-\sum_{i=1}^m f_i(z)\ln(z_i/a_i)+g(z)\geqq 0\Leftrightarrow\sum_{i=1}^m h_i'(z_i)f_i(z)\leqq g(z)\Rightarrow\sum_{i=1}^m h_i'(z_i)f_i(z)\leqq MH(z)+M(\exp(1)-1)\sum_{i=1}^m a_i+N$ for all $z\in P^m$. That is, H satisfies (H1)-(H3) and (H6). Consequently, (H1)-(H3) and (H6) are a generalization of Groger's dissipativity condition. Hence, from the theory of Horn, Feinberg, and Jackson [5], [11], [12], on mass-action kinetics, it follows that many systems of practical interest satisfy these assumptions. Also, the a priori bounds obtained in § 3 apply to these systems.

There are also several model systems that satisfy (H1)-(H6). Before giving a few of these, we state the following theorem, whose proof can be found in [15].

THEOREM 5.1. *If for all* $1\leqq i\leqq m$, f_i *satisfies* $f_i(z)\geqq 0$ *whenever* $z\in\mathrm{cl}(\mathbf{P}^m)$ *and* $z_i=0$ *and* (A3) *holds with* $\mathbf{I}=\mathrm{cl}(P^m)$, *then* (2.1) *is invariant on* $\mathrm{cl}(\mathbf{P}^m)$.

5.1. Model systems. (I) The Brusselator is described in Prigogine and Nicolis [18] and models a simple reaction-diffusion system. The equations are given by

$$\begin{aligned}
&u_t(x,t)=d_1\Delta u(x,t)+Fv-v^2(x,t)u(x,t), \quad x\in\Omega, \quad t>0,\\
&v_t(x,t)=d_2\Delta v(x,t)+E-(F+1)v(x,t)+v^2(x,t)u(x,t), \quad x\in\Omega, \quad t>0,\\
&B(u(x,t),v(x,t))=\gamma, \quad x\in\partial\Omega, \quad t>0,\\
&(u(x,0),v(x,0))=(u_0(x),v_0(x)), \qquad x\in\Omega,
\end{aligned} \tag{5.2}$$

where d_1, d_2, E, $F>0$, and u_0, $v_0 \in L^\infty(\Omega, [0,\infty))$. If we denote $f_1(u,v) = Kv - v^2u$ and $f_2(u,v) = L-(K+1)v+v^2u$, then clearly Theorems 2.1 and 5.1 are satisfied. Hence, (5.2) is invariant on cl $(\mathbf{P}^2)$, and if $H(u,v) = u+v$, then (H1)-(H6) hold with $A = \left(\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right)$, $r=1$, $K_1 = F$, $K_2 = 0$, $K_5 = F+1$, $K_6 = E$, $q_1 = 3$, $K_7 = 0$, and $K_8 = E$. Thus, from Theorems 3.2 and 2.2, $T_{\max} = \infty$.

We note that Hollis, Martin, and Pierre [10] show u and v to be uniformly bounded independently of α, β, and γ, as long as (A1) and (A3) are satisfied. Still, this example illustrates the ease with which our main results can be applied.

(II) Rothe [19, p. 157] considers the system

$$\begin{cases} u_t(x,t) = d_1\Delta u(x,t) + w(x,t) - u(x,t)v(x,t), \\ v_t(x,t) = d_2\Delta v(x,t) + w(x,t) - u(x,t)v(x,t), & x\in\Omega, \quad t>0, \\ w_t(x,t) = d_3\Delta w(x,t) + u(x,t)v(x,t) - w(x,t), \\ \begin{cases} B(u(x,t), v(x,t)) = \gamma, & x\in\partial\Omega, \quad t>0, \\ (u(x,0), v(x,0), w(x,0)) = (u_0(x), v_0(x), w_0(x)), & x\in\Omega \end{cases} \end{cases} \tag{5.3}$$

as a model for the reaction

$$U + V \rightleftarrows W, \tag{5.4}$$

where u, v, and w are the concentrations of U, V, and W, respectively, reacting according to (5.4) with diffusion rates d_1, d_2, $d_3>0$, and initial concentrations u_0, v_0, $w_0 \in L^\infty(\Omega, [0,\infty))$. For the case of Neumann boundary conditions, Rothe [19] shows that $T_{\max} = \infty$ if $n \leqq 5$.

Suppose n is arbitrary, $f_1(u,v,w) = f_2(u,v,w) = w - uv$, and $f_3(u,v,w) = uv - w$. Then Theorems 2.1 and 5.1 are satisfied, and (5.3) is invariant on cl $(\mathbf{P}^3)$. Also, if $H(u,v,w) = u+v+2w$, then (H1)-(H6) hold with

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix},$$

$r=1$, $K_1=1$, $K_2=0$, $K_5=1$, $K_6=0$, $q_1=2$, and $K_7 = K_8 = 0$. Thus, from Theorems 3.2 and 2.2 we have $T_{\max} = \infty$.

(III) Lasry [14] considers the following system as a model to study the properties of nerve conduction:

$$\begin{aligned} &u_t(x,t) = d_1\Delta u(x,t) + K(1-\rho(x,t))u(x,t) - g(\beta(x,t))v(x,t), && x\in\Omega, \quad t>0, \\ &v_t(x,t) = d_2\Delta v(x,t) + g(\beta(x,t))u(x,t) + K(1-\rho(x,t))v(x,t), && x\in\Omega, \quad t>0, \\ &B(u(x,t), v(x,t)) = \gamma, && x\in\partial\Omega, \quad t>0, \\ &(u(x,0), v(x,0)) = (u_0(x), v_0(x)), && x\in\Omega. \end{aligned} \tag{5.5}$$

Here, d_1, $d_2>0$, u_0, $v_0 \in L^\infty(\Omega, R)$, (ρ, β) are polar coordinates for (u,v), $K>0$, and g is a smooth 2π periodic function. Note that although β and hence g are undefined when $\rho = 0$, ug and vg can be defined to be zero when $u = v = 0$. Hence, if $f_1(u,v) = K(1-\rho)u - g(\beta)v$ and $f_2(u,v) = g(\beta)u + K(1-\rho)v$, then f_1 and f_2 are locally Lipschitz. Consequently, Theorem 2.1 holds, and if $H(u,v) = u^2+v^2$ for all $(u,v)\in P^2$, then (H1)-(H6) are satisfied with $A = \left(\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right)$, $r=1$, $K_1 = 4(K + |g|_{\infty,[0,2\pi]})$, $K_2 = 0$, $K_5 = 8(K+|g|_{\infty,[0,2\pi]})$, $q_1 = 3$, $K_6 = 0$, $K_7 = K$, and $K_8 = 0$. Hence, from Theorems 3.2 and 2.2, we have $T_{\max} = \infty$.

REFERENCES

[1] N. D. ALIKAKOS, *L_p-bounds of solutions of reaction-diffusion equations*, Comm. Partial Differential Equations, 4 (1979), pp. 827-868.

[2] H. AMANN, *Global existence for semilinear parabolic systems*, J. Reine Angew. Math., 360 (1985), pp. 47-83.

[3] J. F. G. AUCHMUTY AND G. NICOLIS, *Bifurcation analysis of nonlinear reaction-diffusion equations—I. Evolution equations and steady state solutions*, Bull. Math. Biol., 37 (1975), pp. 323-365.

[4] P. W. BATES AND K. J. BROWN, *Convergence to equilibrium in a reaction-diffusion system*, Nonlinear Anal. Theory Methods Appl., 8 (1984), pp. 227-235.

[5] M. FEINBERG, *Complex balancing in general kinetic systems*, Arch. Rational Mech. Anal., 49 (1972), pp. 187-194.

[6] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.

[7] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[8] K. GROGER, *On the existence of steady states of certain reaction-diffusion systems*, Arch. Rational Mech. Anal., 1 (1986), pp. 297-306.

[9] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1981.

[10] S. HOLLIS, R. MARTIN, AND M. PIERRE, *Global existence and boundedness in reaction-diffusion systems*, SIAM J. Math. Anal., 18 (1987), pp. 744-761.

[11] F. HORN AND R. JACKSON, *General mass action kinetics*, Arch. Rational Mech. Anal., 47 (1972), pp. 81-116.

[12] F. HORN, *Necessary and sufficient conditions for complex balancing in chemical kinetics*, Arch. Rational Mech. Anal., 49 (1972), pp. 172-186.

[13] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URALCEVA, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Transl. 33, American Mathematical Society, Providence, RI, 1968.

[14] J. M. LASRY, International working paper of the Mathematical Research Center, Cérémade, University of Paris-Dauphine, Paris, 1975.

[15] J. H. LIGHTBOURNE AND R. H. MARTIN, *Relatively continuous nonlinear perturbations of analytic semigroups*, J. Nonlinear Anal. Theory Methods Appl., 1 (1977), pp. 277-292.

[16] K. MASUDA, *On the global existence and asymptotic behavior of solutions of reaction-diffusion equations*, Hokkaido Math. J., 12 (1982), pp. 360-370.

[17] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, Berlin, New York, 1983.

[18] J. PRIGOGINE AND G. NICOLIS, *Self-Organization in Nonequilibrium Systems*, Wiley-Interscience, New York, 1977.

[19] F. ROTHE, *Global Solutions of Reaction-Diffusion Systems*, Lecture Notes in Math. 1072, Springer-Verlag, Berlin, New York, 1984.

[20] J. A. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1983.

[21] R. SPERB, *Maximum Principles and Their Applications*, Math. Sci. Engrg. 157, Academic Press, New York, 1981.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1145-1171, September 1989

008

BIFURCATION OF HOMOCLINIC ORBITS AND BIFURCATION FROM THE ESSENTIAL SPECTRUM*

C. A. STUART†

Abstract. Bifurcation for nonlinear eigenvalue problems involving a second-order ordinary differential equation on the line is considered. Solutions are required to vanish at infinity in both directions and so correspond to homoclinic orbits. When posed in function spaces, the problem concerns bifurcation from the continuous spectrum. The present approach is based on a rescaling that reduces the problem to that of continuing a nontrivial homoclinic orbit in a context where the perturbations are not periodic and are not smooth with respect to uniform convergence. Nonetheless, the nondegeneracy required for continuation amounts to finding simple zeros of a function analogous to Melnikov's function.

Key words. bifurcation, essential spectrum, Melnikov function, homoclinic orbit

AMS(MOS) subject classification. 34B15

1. Introduction.

We consider the following eigenvalue problem:

$$u''(x)+\lambda u(x)+f(\lambda, x, u(x), u'(x))=0 \quad \text{for } x\in\mathbb{R}, \tag{1.1}$$

$$\lim_{x\to-\infty} u(x)=\lim_{x\to+\infty} u(x)=0 \tag{1.2}$$

where the function f satisfies the following conditions:

(A1) $f\in C^1(\mathbb{R}^4)$ and for all $(\lambda, x)\in\mathbb{R}^2$, $f(\lambda, x, ., .)\in C^2(\mathbb{R}^2)$. Furthermore, there exists $B\in C(\mathbb{R}^2)$ such that for all $(\lambda, x, p, q)\in\mathbb{R}^4$,

$$|f(\lambda, x, p, q)|\leqq B(\lambda, p)(p^2+q^2)^{1/2},$$

$$|\partial_i f(\lambda, x, p, q)|\leqq B(\lambda, p)(p^2+q^2)^{1/2} \quad \text{for } i=2, 3, 4.$$

This means that $u\equiv 0$ is a solution of (1.1), (1.2) for every $\lambda\in\mathbb{R}$ and we are interested in pairs (λ, u) satisfying (1.1), (1.2) with $u\not\equiv 0$. This kind of problem has often been used as an example in the study of bifurcation from the essential spectrum [7]-[11], [13]-[15], [17] since in appropriate function spaces the linearisation of (1.1), (1.2) about $u\equiv 0$ has the interval $[0, \infty)$ as spectrum. In this setting, conditions are given determining whether or not there is bifurcation in $L^p(\mathbb{R})$ from $\lambda=0$ in the following sense:

(1.3) There exists a sequence $\{(\lambda_n, u_n)\}$ of solutions of (1.1), (1.2) such that $u_n\not\equiv 0$, $\lambda_n\to 0$, and $|u_n|_p\to 0$ where $|\cdot|_p$ denotes the usual norm on $L^p(\mathbb{R})$.

Although (1.3) constitutes the basic definition of bifurcation at $\lambda=0$, it is possible and even desirable to envisage bifurcation in stronger senses such as follows:

(1.4) There exists a connected set $\mathscr{C}$ in $\mathbb{R}\times[L^p(\mathbb{R})\backslash\{0\}]$ such that (λ, u) is a solution of (1.1), (1.2) for all $(\lambda, u)\in\mathscr{C}$ and the solution $(0, 0)$ belongs to the closure of $\mathscr{C}$ in $\mathbb{R}\times L^p(\mathbb{R})$.

or even:

(1.5) There exist $\delta>0$ and $\phi\in C((-\delta, 0), L^p(\mathbb{R}))$ such that, for $-\delta<\lambda<0$, $(\lambda, \phi(\lambda))$ is a solution of (1.1), (1.2) with $\phi(\lambda)\not\equiv 0$ and $\lim_{\lambda\to 0-}|\phi(\lambda)|_p=0$.

* Received by the editors September 28, 1987; accepted for publication (in revised form) October 25, 1988.

† Département de Mathématique, Ecole Polytechnique Fédérale Lausanne, CH-1015 Lausanne, Switzerland.

Clearly, (1.5) implies (1.4), which in turn implies (1.3); but in general these notions of bifurcation are distinct.

Bifurcation in the sense of (1.3) can be established by variational methods [10]–[13], [16]. Very roughly, if f has the special form

$$f(\lambda, x, p, q) = r(x)|p|^{\sigma}p \tag{1.6}$$

where $r \in C(\mathbb{R})$ and $\sigma > 0$, the variational approach has been applied to cases where either $Q = \lim_{|x|\to\infty} r(x) = 0$ or

$$Q = \lim_{|x|\to\infty} r(x) > 0 \quad \text{and} \quad \tfrac{1}{2}\{r(x)+r(-x)\} \gneqq Q \quad \forall x \in \mathbb{R}$$

(see [12], [13], [15], [16] for other conditions and generalisations), and the solutions (λ_n, u_n) obtained in this way have a simple variational characterisation as follows. u_n minimizes an appropriate potential function over a manifold of codimension 1. However, the variational approach does not yield information about bifurcation in the sense of (1.4) or (1.5) and, of course, not all equations have a variational structure.

Bifurcation in the sense of (1.4) has been discussed by Toland [17] for the special case (1.6). His method replaces (1.1), (1.2) by the same equation on the interval $(-L, L)$ with boundary conditions $u(-L) = u(L) = 0$. To extract useful information about the limit $L \to \infty$, Toland's method must require that $r(x) = r(-x)$ for all $x \in \mathbb{R}$ and r nonincreasing on $[0, \infty)$. Other results on this kind of bifurcation are contained in [1] and [18].

Bifurcation in the sense of (1.5) was first discussed by Küpper and Reimer [7] in the case (1.6) where r is constant on $\mathbb{R}$. In this case (1.1) is autonomous and can be analysed by quadrature. For nonconstant r, the problem is treated in [14], where in the context of (1.6) it is required that

$$r(x) = r(-x) \quad \forall x \in \mathbb{R} \text{ and } Q \equiv \lim_{|x|\to\infty} r(x) > 0.$$

More recently, Magnus [8], [9] has developed a powerful apparatus for establishing bifurcation in the sense of (1.5) and has applied it to some specific examples of the type (1.6).

In this note, we continue the study of bifurcation in the sense of (1.5) for the problem (1.1), (1.2). The present approach is a little different from that used by Magnus [9], but the two methods have much in common. It may be said that Magnus [9] reduces (1.1), (1.2) to the problem of perturbing a nondegenerate critical point of an appropriate potential function, whereas here we reduce (1.1), (1.2) to a situation in which bifurcation can be established by an application of the Crandall–Rabinowitz theorem [5] on bifurcation from a simple eigenvalue. Both treatments require several changes of variable, but in the present approach they seem more intuitive and consequently are easier to find in some systematic way. Furthermore, our discussion brings out the relationship between our conditions and work on bifurcations of homoclinic orbits using Melnikov's method [2], [3], [6]. Note that (1.1), (1.2) corresponds to seeking homoclinic orbits of (1.1). Finally we note that our approach does not require any variational structure and the behaviour of $|\cdot|_p$ as $\lambda \to 0$ is determined.

After some preliminary work in § 2, the main result is Theorem 3.2 of § 3. However, the hypotheses of that theorem are not formulated directly in terms of the function f in (1.1). Checking these hypotheses involves choosing new variables in ways that depend on the behaviour of f. The remainder of the paper is devoted to showing how this can be done in various circumstances. First, the special form (1.6) is dealt with in § 4, then various more complicated cases are treated in § 5 as higher-order perturbations of (1.6).

2. Rescaling. As in [14], the first step is to introduce the new variables

$$k = |\lambda|^{1/2} \quad \text{and} \quad v(x) = k^{-\alpha} u\left(\frac{x}{k}\right) \tag{2.1}$$

where $\alpha > 0$ is a constant to be chosen later. Then, if (k, v) satisfies

$$v''(x) - v(x) + k^{-(2+\alpha)} f\left(-k^2, \frac{x}{k}, k^{\alpha} v(x), k^{\alpha+1} v'(x)\right) = 0 \quad \forall x \in \mathbb{R}, \tag{2.2}$$

$$\lim_{x \to -\infty} v(x) = \lim_{x \to +\infty} v(x) = 0, \tag{2.3}$$

$$k > 0, \tag{2.4}$$

it is easily verified that (λ, u) satisfies (1.1), (1.2), where

$$\lambda = -k^2 \quad \text{and} \quad u(x) = k^{\alpha} v(kx). \tag{2.5}$$

Furthermore,

$$|u|_p = |\lambda|^{(\alpha - 1/p)/2} |v|_p. \tag{2.6}$$

Since we are trying to solve (1.1), (1.2) near $\lambda = 0$, we suppose that $\alpha > 0$ can be chosen in such a way that (2.2) has a well-defined but nontrivial limit as $k \to 0+$.

(A2) There exists $\alpha > 0$ and $g \in C^2(\mathbb{R}^2)$ such that

$$\lim_{k \to 0+} k^{-(2+\alpha)} f\left(-k^2, \frac{x}{k}, k^{\alpha} p, k^{\alpha+1} q\right) = g(p, q)$$

for all $x \neq 0$ and all $p, q \in \mathbb{R}$. Furthermore, there exists $B \in C(\mathbb{R})$ such that for all $(p, q) \in \mathbb{R}^2$,

$$|g(p, q)| \leqq B(p)(p^2 + q^2)^{1/2},$$

$$|\partial_i g(p, q)| \leqq B(p)(p^2 + q^2)^{1/2} \quad \text{for } i = 1, 2.$$

When (A2) is satisfied, we set

$$h(k, x, p, q) = k^{-(2+\alpha)} f\left(-k^2, \frac{x}{k}, k^{\alpha} p, k^{\alpha+1} q\right) - g(p, q) \tag{2.7}$$

and note that in general $h(k, o, p, q)$ does not tend to zero as $k \to 0+$. Due to this nonuniformity of the limit in (A2), the standard results and methods concerning the bifurcation of homoclinic orbits cannot be applied to (2.2), (2.3) as $k \to 0$. The limit must be taken in some space such as $L^p(\mathbb{R})$ or $H^{-1}(\mathbb{R})$ that allows for nonuniform convergence. Once this space has been chosen, the domain of the operators concerned is fixed by the requirement that $u'' - u$ should define an isomorphism between the spaces concerned. For simplicity, in what follows we only discuss the case

$$X = H^1(\mathbb{R}) \quad \text{and} \quad Y = H^{-1}(\mathbb{R}), \tag{2.8}$$

but other frameworks such as

$$X = W^{2,1}(\mathbb{R}) \quad \text{and} \quad Y = L^1(\mathbb{R}) \tag{2.9}$$

are possible and would yield results under different conditions on the nonlinear term f. Also, a dynamical-systems approach does not control $|\cdot|_p$.

From now on we fix the setting as (2.8) and we must ensure a certain minimal regularity of the operator induced by f between these spaces. To obtain this regularity

it is usually necessary to rescale the parameter k. For this, we denote by Γ a function having the following properties:

$$\Gamma \in C^1((0, \varepsilon), (0, \infty)) \quad \text{with } \lim_{k \to 0+} \Gamma(k) = 0 \text{ and } \Gamma'(k) > 0 \quad \text{for } 0 < k < \varepsilon. \tag{2.10}$$

In this case the inverse function Γ^{-1} will be denoted by γ, and we introduce a new variable s related to k by

$$s = \Gamma(k), \qquad k = \gamma(s). \tag{2.11}$$

Recalling that $X = H^1(\mathbb{R})$ and $Y = H^{-1}(\mathbb{R}) = X^*$, we note that an isomorphism $L: X \to Y$ is defined by

$$\langle Lv, \phi \rangle = -\int_{-\infty}^{\infty} v'\phi' + v\phi \, dx \tag{2.12}$$

where $\langle ., . \rangle$ denotes the duality between X and X^*.

Furthermore, by the growth conditions in (A1) and (A2), operators $G: X \to Y$ and $H: (0, \Gamma(\varepsilon)) \times X \to Y$ can be defined by

$$\langle G(v), \phi \rangle = \int_{-\infty}^{\infty} g(v(x), v'(x))\phi(x) \, dx, \tag{2.13}$$

$$\langle H(s, v), \phi \rangle = \int_{-\infty}^{\infty} h(\gamma(s), x, v(x), v'(x))\phi(x) \, dx \tag{2.14}$$

for $v, \phi \in X$, since $g(v(\ ), v'(\cdot))$ and $h(\gamma(s), \cdot, v(\ ), v'(\ )) \in L^2(\mathbb{R})$ for all $v \in X$.

Setting $J = (-\Gamma(\varepsilon), \Gamma(\varepsilon))$ we extend H to $J \times X$ by setting $H(0, v) = 0$ for all $v \in X$ and $H(s, v) = H(-s, v)$ for $-\Gamma(\varepsilon) < s < 0$ and $v \in X$.

Finally, we define $F: J \times X \to Y$ by

$$F(s, v) = Lv + G(v) + H(s, v). \tag{2.15}$$

Clearly, if $F(s, v) = 0$ then $(\gamma(s), v)$ satisfies (2.2), (2.3) in the weak sense provided that $s \neq 0$. However, the conditions (A1) and (A2) ensure that v is actually a classical solution decaying exponentially to zero as $|x| \to \infty$.

LEMMA 2.1. *Let the conditions* (A1) *and* (A2) *be satisfied and let*

$$F(s, v) = 0 \quad \textit{for } (s, v) \in J \times X \quad \text{with } v \neq 0. \tag{2.16}$$

Then

$$v \in C^3(\mathbb{R}) \cap H^3(\mathbb{R}), \tag{2.17}$$

$$\lim_{|x| \to \infty} |x|^{-1} \log \{v(x)^2 + v'(x)^2\}^{1/2} = -1. \tag{2.18}$$

If $s > 0$, *then* $(\gamma(s), v)$ *satisfies* (2.2), (2.3), (2.4) *classically, whereas if* $s = 0$, *then* v *satisfies*

$$v''(x) - v(x) + g(v(x), v'(x)) = 0 \quad \forall x \in \mathbb{R}. \tag{2.19}$$

Remarks. By (2.17), v' and $v'' \in X$ when (1.16) is satisfied. The result (2.18) implies that for all $\varepsilon \in (0, 1)$

$$\lim_{|x| \to \infty} e^{(1-\varepsilon)|x|} v(x) = \lim_{|x| \to \infty} e^{(1-\varepsilon)|x|} v'(x) = 0. \tag{2.20}$$

From (2.20) and results on asymptotic equivalence we can deduce that (2.20) holds even for $\varepsilon = 0$, but since we do not need this result we omit it.

Proof. From (2.16) it follows that

$$\int_{-\infty}^{\infty} v'\phi'\,dx = \int_{-\infty}^{\infty} \{-v+G(v)+H(s,v)\}\phi\,dx \quad \forall \phi \in X$$

and by (A1) and (A2), $-v+G(v)+H(s,v)\in L^2(\mathbb{R})$ for all $(s,v)\in J\times X$. Hence $v\in H^2(\mathbb{R})$ and $v''-v+G(v)+H(s,v)=0$ almost everywhere on $\mathbb{R}$. But $v\in H^2(\mathbb{R})$ implies that $v\in C^1(\mathbb{R})$ with

$$\lim_{|x|\to\infty} v(x) = \lim_{|x|\to\infty} v'(x) = 0. \tag{2.21}$$

Thus by (A1) and (A2), $-v+G(v)+H(s,v)\in C(\mathbb{R})$ and so $v\in C^2(\mathbb{R})$. This proves that v satisfies either (2.2) when $s\neq 0$ or (2.19) when $s=0$ in the classical sense. From the smoothness of f and g given by (A1) and (A2) it then follows that $v\in C^3(\mathbb{R})$. By differentiating (2.2) (respectively, (2.19)) with respect to x and using (2.21) we can conclude that $v\in H^3(\mathbb{R})$. Finally, (2.21) and the estimates for f and g given by (A1) and (A2) mean that (2.18) can be deduced from Theorem 5 of [4, Chap. IV] applied to (2.2) (respectively, (2.19)).

In view of this result, bifurcation in the sense of (1.5) can be established for (1.1), (1.2) by showing the following:

(2.22) There exist $\delta>0$ and $\psi\in C((0,\delta), X\setminus\{0\})$ such that $F(s,\psi(s))=0$ for $0<s<\delta$.

In fact the pair $(\lambda,\phi(\lambda))$ defined by

$$\lambda = -\gamma(s)^2, \qquad \phi(\lambda)(x) = |\lambda|^{\alpha/2}\psi(\Gamma(|\lambda|^{1/2}))(|\lambda|^{1/2}x) \tag{2.23}$$

satisfies (1.1), (1.2) for $0<s<\delta$, $s=\Gamma(\sqrt{-\lambda})$ and $\phi\in C((-\delta_1,0), X\setminus\{0\})$. Hence there is bifurcation for (1.1), (1.2) in the sense of (1.5) provided that

$$\lim_{\lambda\to 0-} |\phi(\lambda)|_p = \lim_{\lambda\to 0-} |\lambda|^{(\alpha-1/p)/2}|\psi(\Gamma(|\lambda|^{1/2}))|_p = 0. \tag{2.24}$$

One way of resolving (2.22) and (2.24) simultaneously is to show that

(2.25) There exist $\delta>0$ and $\psi\in C([0,\delta), X\setminus\{0\})$ such that $F(s,\psi(s))=0$ for $0\leqq s<\delta$.

If (2.25) is satisfied it follows that there is bifurcation in the sense of (1.5) whenever $p\geqq \max\{2, 1/\alpha\}$ because

$$\lim_{\lambda\to 0-} |\psi(\Gamma(|\lambda|^{1/2}))|_p = |\psi(0)|_p > 0 \quad \text{since } X = H^1(\mathbb{R})$$

is continuously embedded in $L^p(\mathbb{R})$ for $2\leqq p\leqq\infty$.

In the next section we give conditions implying that (2.25) is satisfied. The method used requires a certain degree of smoothness of the function $F: J\times X\to Y$ that is ensured by the following assumption:

(A3) $F\in C^1(J\times X, Y)$ and the second-order derivatives D_v^2F, D_sD_vF, and D_vD_sF exist and are continuous on $J\times X$.

By strengthening (A1) and (A2), we can ensure (as in §§ 4 and 5) that $F(s,\cdot)\in C^2(X,Y)$ for each $s\in J$. Thus the content of (A3) really concerns the smoothness of F with respect to s, and the point is to be able to choose Γ and γ given by (2.10), (2.11) so that (A3) is valid.

3. Continuation of a homoclinic orbit. In the previous section the problem of bifurcation in the sense of (1.5) for (1.1), (1.2) has been reduced to (2.25). One way of establishing (2.25) is to suppose that the limit equation $F(0, v) = 0$ has a nontrivial solution v_0 in X and then to ensure that v_0 can be continued to yield a branch of solutions of $F(s, v) = 0$ for s near 0. Thus we begin with the following assumption:

(A4) There exists $v_0 \in X \setminus \{0\}$ such that $F(0, v_0) = 0$.

Next we must establish properties of v_0 and F that will allow us to continue v_0. First we note that the equation $F(0, v_0) = 0$ is equivalent to the autonomous equation (2.19) with v_0 satisfying (2.17) and (2.18). Setting $z = v_0'$, we have that $z \in X$ and satisfies

$$z''(x) - z(x) + a(x)z(x) + b(x)z'(x) = 0 \quad \forall x \in \mathbb{R} \tag{3.1}$$

where

$$a(x) = \partial_1 g(v_0(x), v_0'(x)) \quad \text{and} \quad b(x) = \partial_2 g(v_0(x), v_0'(x)). \tag{3.2}$$

From (2.20) and (A2) it follows that a and b decay exponentially to zero as $|x| \to \infty$. Hence the integrating factor for (3.1) defined by

$$i(x) = \exp \int_0^x b(y)\, dy \tag{3.3}$$

has the following properties:

$$i \in C^2(\mathbb{R}), \quad i' = bi \quad \text{and} \quad 0 < L_1 \leqq i(x) \leqq L_2 < \infty \quad \forall x \in \mathbb{R}. \tag{3.4}$$

From this it follows easily that

$$v \in X \Leftrightarrow iv \in X \tag{3.5}$$

and (3.1) is equivalent to its symmetric form

$$(iz')' - i[1 - a]z = 0 \quad \text{on } \mathbb{R}. \tag{3.6}$$

To discuss the properties of the linearisation of F at v_0, it is convenient to define a linear operator $C(v_0)\colon X \to Y$ by

$$C(v_0)u = au + bu' \tag{3.7}$$

where a and b are given by (3.2). Then by (A1)-(A3),

$$D_v F(0, v_0)u = Lu + C(v_0)u \quad \forall u \in X. \tag{3.8}$$

When no confusion is likely to arise we write C for $C(v_0)$.

Recalling (2.8), we have the following version of the Fredholm alternative.

LEMMA 3.1. *Let the conditions* (A1)-(A4) *be satisfied. Then*

(i) $C\colon X \to Y$ *is compact*;

(ii) $\ker(L + C) = \operatorname{span}\{v_0'\}$;

(iii) $\langle (L + C)w, iu \rangle = \langle (L + C)u, iw \rangle$, for all $u, w \in X$;

(iv) *For* $\xi \in Y$, *the equation* $(L + C)u = \xi$ *has a solution in* X *if and only if* $\langle \xi, iv_0' \rangle = 0$.

Proof. (i) This is an easy consequence of the fact that $\lim_{|x| \to \infty} a(x) = \lim_{|x| \to \infty} b(x) = 0$ together with the compactness of the Sobolev embeddings on compact intervals.

(ii) If $u \in \ker(L + C)$, we have that $u \in X$ and $\langle (L + C)u, \phi \rangle = 0$, for all $\phi \in X$. In particular, u is a weak solution of the linear equation (3.1) having coefficients a and b in $C^1(\mathbb{R})$. Thus, $u \in C^3(\mathbb{R})$ and satisfies (3.1) classically. From (3.1) we can now conclude that $u \in H^2(\mathbb{R})$ and so, in particular,

$$\lim_{|x| \to \infty} u(x) = \lim_{|x| \to \infty} u'(x) = 0. \tag{3.9}$$

But v_0' also satisfies (3.1) and the conditions (3.9). Hence, if $u \notin \operatorname{span}\{v_0'\}$ it would follow that all solutions of (3.1) satisfy (3.9). Since a and b decay exponentially to zero as $|x| \to \infty$, this would contradict Theorem 2 of [4, Chap. IV]. Hence we may conclude that $u \in \operatorname{span}\{v_0'\}$.

(iii) For $u, w \in X$,

$$\begin{aligned}\langle (L+C)w, iu\rangle &= -\langle w', (iu)'\rangle + \langle -w + aw + bw', iu\rangle \\ &= -\langle w', biu + iu'\rangle + \langle -w + aw + bw', iu\rangle \\ &= -\langle w', iu'\rangle + \langle -w + aw, iu\rangle \\ &= \langle (L+C)u, iw\rangle.\end{aligned}$$

(iv) Since $L: X \to Y$ is an isomorphism and $C: X \to Y$ is compact, the Fredholm alternative implies that $(L+C)u = \xi$ has a solution if and only if $\langle \xi, v\rangle = 0$ for all $v \in \ker (L+C)^*$. But $v \in \ker(L+C)^*$ means that $v \in Y^*$ and $\langle (L+C)^* v, w\rangle = 0$, for all $w \in X$. Since $Y^* = X$, this is equivalent to the conditions

$$v \in X \quad \text{and} \quad \langle (L+C)w, v\rangle = 0 \quad \forall w \in X.$$

By (3.5), this can be expressed equivalently as

$$v = iz \quad \text{where } z \in X \quad \text{and} \quad \langle (L+C)w, iz\rangle = 0 \quad \forall w \in X,$$

and using part (iii) this becomes

$$v = iz \quad \text{where } z \in X \quad \text{and} \quad \langle (L+C)z, iw\rangle = 0 \quad \forall w \in X.$$

When we appeal once again to (3.5) this is equivalent to

$$v = iz \quad \text{and} \quad z \in \ker (L+C)$$

and so by part (ii), it follows that

$$\ker (L+C)^* = i \ker (L+C) = i \operatorname{span}\{v_0'\}.$$

This proves (iv).

We now turn to the problem of continuing the solution v_0 given by (A4). The simplest approach would be to show that $D_v F(0, v_0): X \to Y$ is an isomorphism and hence to apply the Implicit Function Theorem. However, this cannot be done since $D_v F(0, v_0)$ is not invertible on X and v_0 is not an isolated zero of $F(0, \cdot)$ in X. In fact, defining a translation operator T_τ by

$$T_\tau v(x) = v(x+\tau) \quad \text{for } x, \tau \in \mathbb{R} \tag{3.10}$$

we see that

$$F(0, T_\tau v_0) = 0 \quad \forall \tau \in \mathbb{R}. \tag{3.11}$$

Furthermore, $\tau \to T_\tau v_0$ is continuously differentiable since

$$v_0' \in X \quad \text{and} \quad \frac{d}{dt}(T_\tau v_0) = T_\tau v_0'.$$

Hence, $0 = (d/dt)F(0, T_\tau v_0) = D_v F(0, T_\tau v_0) T_\tau v_0'$ for all $\tau \in \mathbb{R}$ and in particular, as we already know,

$$v_0' \in \ker D_v F(0, v_0) = \ker (L+C).$$

One way of resolving this difficulty is to suppose that f has symmetries (even with respect to x and $u'(x)$) allowing us to remove the above degeneracy by restricting the discussion to the subspaces of X and Y consisting of even elements. This is what is done in [14] but in the present discussion we are trying to obtain results for cases where f does not have these symmetries. The idea now is to consider the curve

$$\mathscr{C} = \{(0, T_\tau v_0) : \tau \in \mathbb{R}\} \subset \mathbb{R} \times X \tag{3.12}$$

as a curve of trivial solutions for the equation $F(s, v) = 0$ and to seek values of the phase parameter τ at which a branch of solutions bifurcates from $\mathscr{C}$ into the region $(0, \infty) \times X$.

We shall show that the appropriate values of τ are characterised as the simple zeros of the function $M: \mathbb{R} \to \mathbb{R}$ defined by

$$M(\tau) = \langle D_s H(0, T_\tau v_0), T_\tau(iv_0') \rangle. \tag{3.13}$$

We note that since $v_0'' \in X$, it follows from (A3) that $M \in C^1(\mathbb{R})$ with

$$M'(\tau) = \langle D_u D_s H(0, T_\tau v_0) T_\tau v_0', T_\tau(iv_0') \rangle + \langle D_s H(0, T_\tau v_0), T_\tau(iv_0')' \rangle. \tag{3.14}$$

The function M defined by (3.13) depends on the choice of the solution v_0 of $F(0, v) = 0$, so we denote it temporarily by $M_{v_0}(\tau)$. For $t \in \mathbb{R}$, $T_t v_0$ is also a solution of $F(0, v) = 0$, so we could use it to define another function $M_{T_t v_0}(\tau)$. However, these two functions are related by

$$M_{v_0}(\tau) = C(t) M_{T_t v_0}(\tau - t) \tag{3.15}$$

where $C(t) = \exp\{\int_0^t \partial_2 g(v_0(y), v_0'(y))\, dy\}$.

When the conditions (A1)-(A4) are satisfied, the function M defined by (3.13) will be referred to as a Melnikov function for (1.1), (1.2). This is reasonable terminology because if our procedure is applied to a situation covered by Melnikov theory [2], [3], [6], then the function M generated by our method coincides with the standard Melnikov function. However, because of the nonuniformity of the limit in (A2), Melnikov theory in its usual form [2], [3], [6], cannot be applied to (2.2), (2.3), (2.4). We prove our result by an application of the classical theorem on bifurcation from a simple eigenvalue in the form due to Crandall and Rabinowitz. Many situations covered by standard Melnikov theory could also be treated in this way.

THEOREM 3.2. *Let the conditions* (A1)-(A4) *be satisfied and suppose that there exists* $\tau_0 \in \mathbb{R}$ *such that*

$$M(\tau_0) = 0 \quad \text{and} \quad M'(\tau_0) \neq 0$$

where $M: \mathbb{R} \to \mathbb{R}$ *is the Melnikov function defined by* (3.13) *using a solution* $v_0 \in X$ *of* $F(0, v) = 0$ *given by* (A4). *Let* $W = \{v \in X : \langle v, T_{\tau_0} v_0' \rangle = 0\}$. *Then*:

(a) *The equation*

$$[L + C(T_{\tau_0} v_0)] z = -D_s H(0, T_{\tau_0} v_0)$$

has a unique solution $z_0 \in W$ *where* C *is defined by* (3.7) *with* v_0 *replaced by* $T_{\tau_0} v_0$. *Let* $Z = \{v \in W : \langle v, z_0 \rangle = 0\}$. *Then*:

(b) *There exist* $\delta > 0$ *and functions*

$$\eta \in C((-\delta, \delta), \mathbb{R}) \quad \text{and} \quad \mu \in C((-\delta, \delta), Z)$$

such that $\eta(0) = \tau_0$, $\mu(0) = 0$, *and* $F(s, \psi(s)) = 0$ *for all* $s \in (-\delta, \delta)$, *where* $\psi(s) = T_{\eta(s)} v_0 + s z_0 + s\mu(s)$.

(c) *Setting* $\delta_1 = \gamma(\delta)^2$ *and*

$$\phi(\lambda)(x) = |\lambda|^{\alpha/2} \psi(\Gamma(|\lambda|^{1/2}))(|\lambda|^{1/2} x) \quad \text{for } -\delta_1 < \lambda < 0,$$

we have that $\phi \in C((-\delta_1, 0), X\backslash\{0\})$ *and* $|\phi(\lambda)|_p = |\lambda|^{(\alpha-1/p)/2}|\psi(\Gamma(|\lambda|^{1/2}))|_p$ *for* $1 \leqq p \leqq \infty$, *where* $\alpha > 0$ *is the constant in* (A2), *where* Γ *and* γ *are the functions given by* (2.10) *and* (2.11) *used to define* H *in* (2.14). *There is bifurcation in* $L^p(\mathbb{R})$ *for* (1.1), (1.2) *in the sense of* (1.5) *provided that* $p \geqq \min\{2, 1/\alpha\}$.

Proof. Replacing v_0 by $T_{\tau_0}v_0$ and recalling (3.15), we suppose henceforth that $\tau_0 = 0$ and so $W = \{w \in X: \langle v_0', w\rangle = 0\}$. Define an operator $N: \mathbb{R}\times J \times W \to Y$ by

$$N(\tau, s, w) = F(s, w + T_\tau v_0)$$

where F is given by (2.15). Thus $\mathscr{C}$ becomes the line $\mathbb{R}\times\{0\}\times\{0\}$ of trivial solutions for N since by (3.11), $N(\tau, 0, 0) = 0$ for all $\tau \in \mathbb{R}$.

By (A3), $N \in C^1(\mathbb{R}\times J \times W, Y)$ and furthermore,

$$D_\tau N(\tau, s, w) = D_u F(s, w + T_\tau v_0) T_\tau v_0',$$

and for $(\mu, z) \in \mathbb{R}\times W$,

$$\begin{aligned} D_{(s,w)}N(\tau, s, w)(\mu, z) &= \mu D_s F(s, w + T_\tau v_0) + D_u F(s, w + T_\tau v_0) z \\ &= \mu D_s H(s, w + T_\tau v_0) + D_u F(s, w + T_\tau v_0) z, \end{aligned}$$

$$D_\tau\{D_{(s,w)}N(\tau, s, w)(\mu, z)\} = \mu D_u D_s H(s, w + T_\tau v_0) T_\tau v_0' + D_u^2 F(s, w + T_\tau v_0)[T_\tau v_0', z].$$

Thus by (A3), N has the regularity required by the Crandall-Rabinowitz theorem [5] and so to establish bifurcation at $\tau_0 = 0$ we must check the following:

(i) $\ker D_{(s,w)}N(0, 0, 0) = \operatorname{span}\{(1, z_0)\}$ where $z_0 \in W$;

(ii) $\operatorname{codim} D_{(s,w)}N(0, 0, 0) = 1$;

(iii) $D_\tau\{D_{(s,w)}N(0, 0, 0)(1, z_0)\} \notin \operatorname{Im} D_{(s,w)}N(0, 0, 0)$.

By insisting that the eigenvector has the form $(1, z_0)$ we ensure that the branch of solutions of $N(\tau, s, w) = 0$ bifurcating at $(0, 0, 0)$ can be parameterized by s.

For (i) we note that $(\mu, z) \in \ker D_{(s,w)}N(0, 0, 0) \Leftrightarrow (\mu, z) \in \mathbb{R}\times W$ and

$$\mu D_s H(0, v_0) + (L + C)z = 0 \tag{3.16}$$

where $C: X \to Y$ is defined by (3.7) with $C = C(v_0)$. Thus, by Lemma 3.1, there is an eigenvector of the form $(1, z_0) \in \mathbb{R}\times X$ if and only if

$$\langle D_s H(0, v_0), iv_0'\rangle = 0, \tag{3.17}$$

and in this case the requirement $z_0 \in W$ determines z_0 uniquely by

$$(L + C)z_0 = -D_s H(0, v_0) \tag{3.18}$$

since $\ker(L + C) = \operatorname{span}\{v_0'\}$. The condition (3.17) is satisfied since $M(0) = 0$. Thus (i) is verified and so is part (a) of the theorem.

For (ii), we note that $\xi \in \operatorname{Im} D_{(s,w)}N(0, 0, 0) \Leftrightarrow \xi \in Y$ and there exists $(\mu, z) \in \mathbb{R}\times W$ such that

$$\xi = \mu D_s H(0, v_0) + (L + C)z.$$

By Lemma 3.1 and (3.17) this is equivalent to $\langle \xi, iv_0'\rangle = 0$, and so

$$\operatorname{Im} D_{(s,w)}N(0, 0, 0) = \{\xi \in Y: \langle \xi, iv_0'\rangle = 0\}.$$

This establishes (ii).

For (iii) we note that it is now enough to show that

$$\langle D_\tau\{D_{(s,w)}N(0, 0, 0)(1, z_0)\}, iv_0'\rangle \neq 0$$

where $z_0 \in W$ is determined by (3.18). Since

$$\begin{aligned} D_{(s,w)}N(\tau, 0, 0)(1, z_0) &= D_sH(0, T_\tau v_0) + D_uF(0, T_\tau v_0)z_0 \\ &= D_sH(0, T_\tau v_0) + Lz_0 + D_uG(T_\tau v_0)z_0 \\ &= D_sH(0, T_\tau v_0) + Lz_0 + (T_\tau a)z_0 + (T_\tau b)z_0' \end{aligned}$$

we find that

$$D_\tau\{D_{(s,w)}N(\tau, 0, 0)(1, z_0)\} = D_uD_sH(0, T_\tau v_0)T_\tau v_0' + (T_\tau a)'z_0 + (T_\tau b)'z_0'$$

where the prime stands for differentiation with respect to x. Thus condition (iii) amounts to showing that

$$\langle D_uD_sH(0, v_0)v_0' + a'z_0 + b'z_0', iv_0' \rangle \neq 0.$$

But

$$\begin{aligned} \langle a'z_0 + b'z_0', iv_0' \rangle &= -\langle Cz_0, (iv_0')' \rangle - \langle az_0' + bz_0'', iv_0' \rangle \\ &= -\langle (L+C)z_0, (iv_0')' \rangle + \langle Lz_0, (iv_0')' \rangle - \langle az_0' + bz_0'', iv_0' \rangle \end{aligned}$$

and

$$\begin{aligned} \langle Lz_0, (iv_0')' \rangle &= -\langle z_0', (iv_0')'' \rangle - \langle z_0, (iv_0')' \rangle \\ &= -\langle z_0', (biv_0')' + (iv_0'')' \rangle - \langle z_0, (iv_0')' \rangle \\ &= -\langle z_0', (biv_0')' + i[1-a]v_0' \rangle - \langle z_0, (iv_0')' \rangle \quad \text{by (3.6)} \\ &= -\langle z_0', (biv_0')' - iav_0' \rangle \\ &= \langle z_0'', biv_0' \rangle + \langle z_0', iav_0' \rangle. \end{aligned}$$

Hence

$$\langle a'z_0 + b'z_0', iv_0' \rangle = -\langle (L+C)z_0, (iv_0')' \rangle = \langle D_sH(0, v_0), (iv_0')' \rangle \quad \text{by (3.18).}$$

Thus condition (iii) is reduced to

$$\langle D_uD_sH(0, v_0)v_0', iv_0' \rangle + \langle D_sH(0, v_0), (iv_0')' \rangle \neq 0$$

and by (3.14) this is just the condition $M'(0) \neq 0$. This proves that condition (iii) is satisfied.

When we note that

$$\mathbb{R} \times W = \text{span}\{(1, z_0)\} \oplus [\{0\} \times Z]$$

it follows by Theorem 1.7 of [5] that there exist $\delta > 0$ and functions

$$\eta \in C((-\delta, \delta), \mathbb{R}), \qquad \theta \in C((-\delta, \delta), \{0\} \times Z)$$

such that

$$\eta(0) = 0, \qquad \theta(0) = (0, 0)$$

and

$$N(\eta(s), s(1, z_0) + s\theta(s)) = 0 \quad \text{for } s \in (-\delta, \delta).$$

Setting $\theta(s) = (0, \mu(s))$, we have that $\mu \in C((-\delta, \delta), Z)$, $\mu(0) = 0$, and

$$\begin{aligned} 0 &= N(\eta(s), s, sz_0 + s\mu(s)) \\ &= F(s, T_{\eta(s)}v_0 + sz_0 + s\mu(s)) \quad \text{for } s \in (-\delta, \delta). \end{aligned}$$

This proves (b) in the case $\tau_0 = 0$.

Finally (c) follows directly from (b) as noted in (2.23), (2.24) of § 2.

Remarks. (1) While this paper was being refereed, a revised version of [9] appeared in *Proc. Roy. Soc. Edinburgh Sect.* A, 110 (1988), pp. 1-25. In an addendum, Magnus derived a result (Theorem 6.4) showing how an equation such as that in part (a) above arises in his context.

(2) An essential step in the above proof involves seeking a solution v near $T_{\tau_0}v_0$ in the form

$$v = w + T_\tau v_0 \quad \text{where } w \in W \text{ and } \tau \text{ is near } \tau_0.$$

In fact all v near $T_{\tau_0}v_0$ in X admit such a representation and it is unique. To see this set $f(\tau, r) = \langle T_\tau v_0, T_{\tau_0} v_0' \rangle - r$, and use the Implicit Function Theorem.

4. The main example. In this section we apply the general method to the case where f has the following special form:

$$f(\lambda, x, p, q) = r(x)|p|^\sigma p \quad \text{for } (\lambda, x, p, q) \in \mathbb{R}^4, \quad \text{or} \tag{4.1}$$

$$= r(x)|p|^{\sigma+1} \quad \text{for } (\lambda, x, p, q) \in \mathbb{R}^4 \tag{4.1'}$$

where

(H1) $r \in C^1(\mathbb{R})$ with $\lim_{|x|\to\infty} r(x) = Q > 0$ and $r' \in L^\infty(\mathbb{R})$, where $\sigma > 1$ in case (4.1) and $\sigma \geqq 1$ in case (4.1').

From (H1) it follows that f satisfies the conditions (A1) and (A2) provided that in (A2) we set

$$\alpha = \frac{2}{\sigma},\ g(p, q) = Q|p|^\sigma p, \text{ or } = Q|p|^{\sigma+1} \text{ and } h(k, x, p, q) = R(x/k)|p|^\sigma p \tag{4.2}$$
$$\text{or } = R(x/k)|p|^{\sigma+1}, \text{ where } R(x) = r(x) - Q.$$

In what follows we will usually write out the formulae that are valid for the case (4.1). The corresponding results for the case (4.1') are obtained by a trivial modification. In particular, the essential formulae for the Melnikov function discussed in Theorem 4.3 are identical for (4.1) and (4.1') since the solution v_0 given by (4.26) is positive.

To obtain the regularity of the Nemytskii operators defined by (2.13)-(2.15) required in (A3), we must make an appropriate choice of the function Γ (and hence γ). This depends on the rate at which r converges to Q and the next hypothesis distinguishes a number of situations in which a successful choice of Γ is known.

(H2) With $R(x) = r(x) - Q$ we suppose that one of the following conditions is satisfied:

(i) There exist $\beta \in (0, 1)$, $\nu \geqq 0$, and $L(\pm\infty) \in \mathbb{R}$ such that

$$\lim_{x\to\pm\infty} R'(x)|x|^{\beta+1}[\ln|x|]^{-\nu} = L(\pm\infty); \tag{4.3}$$

(ii) There exist $L(+\infty) \in \mathbb{R}$ and $\psi \in L^1(\mathbb{R})$ such that

$$\lim_{|x|\to\infty} x\psi(x) = 0, \tag{4.4}$$

$$R(x) = -\frac{xL(+\infty)}{1+x^2} + \psi(x) \quad \forall x \in \mathbb{R}; \tag{4.5}$$

(iii) There exist $\beta \in (1, 3/2)$, $\nu \geqq 0$, and $L(\pm\infty) \in \mathbb{R}$ such that (4.3) holds and $\int_{-\infty}^{\infty} R(x)\,dx = 0$.

Remarks. (1) Since both $(L\pm\infty)$ can be zero, the exponents β and ν are not uniquely determined by (4.3). In fact, if the condition is satisfied for exponents β and ν it is also satisfied for β_1 and ν_1 provided that

$$\beta_1 < \beta \quad \text{and} \quad \nu_1 \geqq 0$$

or

$$\beta_1 = \beta \quad \text{and} \quad \nu_1 \geqq \nu.$$

We shall show that the function $\Gamma\colon (0,1)\to\mathbb{R}$ defined by

$$\Gamma(k) = \int_0^k t^{\beta-1}[-\ln t]^{\nu}\, dt \tag{4.6}$$

when r satisfies (H2)(i) or (H2)(iii) with exponents β and ν, or by

$$\Gamma(k) = k \tag{4.7}$$

when r satisfies (H2)(ii), establishes a change of variables for which (A3) is satisfied. Consequently, a Melnikov function M can be defined for any such pair of exponents β and ν. However, there is at most one pair (β, ν) for which the corresponding M is not identically zero.

(2) When (H2)(i) or (H2)(iii) holds, it follows from l'Hôpital's rule that

$$\lim_{x\to\pm\infty} R(x)|x|^{\beta}[\ln|x|]^{-\nu} = \mp\frac{L(\pm\infty)}{\beta} \tag{4.8}$$

whereas (H2)(ii) implies that

$$\lim_{x\to\pm\infty} R(x)|x| = \mp L(+\infty). \tag{4.9}$$

Consequently, if (H2)(i) or (H2)(ii) holds with $L(\pm\infty)\neq 0$, it follows that $R\notin L^1(\mathbb{R})$. On the other hand, if $R\in L^1(\mathbb{R})$ and $\lim_{|x|\to\infty} xR(x)=0$ then (H2)(ii) holds with $L(+\infty)=0$. The case (H2)(iii) enables us to deal with situations where $R\in L^1(\mathbb{R})$ and $\int_{-\infty}^{\infty} R(x)\, dx = 0$. In this case it is useful to set

$$P(x) = \int_{-\infty}^{x} R(t)\, dt \tag{4.10}$$

and to note that by l'Hôpital's rule

$$\lim_{x\to\pm\infty} P(x)|x|^{\beta-1}[\ln|x|]^{-\nu} = \lim_{x\to\pm\infty} \frac{R(x)|x|^{\beta}}{(1-\beta)}[\ln|x|]^{-\nu} = \pm\frac{L(\pm\infty)}{\beta(\beta-1)}. \tag{4.11}$$

From this it follows that

$$\lim_{x\to\pm\infty} \{xR(x)-P(x)\}|x|^{\beta-1}[\ln|x|]^{-\nu} = \frac{L(\pm\infty)}{1-\beta}. \tag{4.12}$$

When r satisfies (H1) and (H2), the function Γ defined by (4.6) or (4.7) satisfies (2.10) and its inverse γ is defined on $(0,\Gamma(1))$. Furthermore, for $0<s<\Gamma(1)$,

$$\gamma'(s)\gamma(s)^{\beta-1}[-\ln\gamma(s)]^{\nu} = 1. \tag{4.13}$$

In particular, when $\nu=0$ we have that

$$\gamma(s) = (\beta s)^{1/\beta}. \tag{4.14}$$

In the present context, the operators G and H defined by (2.13)-(2.15) are given by

$$G(v)(x) = Q|v(x)|^{\sigma}v(x), \tag{4.15}$$

$$H(s, v)(x) = \begin{cases} R\left(\dfrac{x}{\gamma(s)}\right)|v(x)|^{\sigma}v(x) & \text{for } 0 < s < \Gamma(1), \\ 0 & \text{for } s = 0, \\ -H(-s, v)(x) & \text{for } -\Gamma(1) < s < 0, \end{cases} \tag{4.16}$$

$$F(s, v) = Lv + G(v) + H(s, v). \tag{4.17}$$

In (4.15), (4.16) we have supposed that f has the form (4.1). When (4.1′) holds we replace $|v|^{\sigma}v$ by $|v|^{\sigma+1}$ in these definitions.

LEMMA 4.1. *Let f satisfy the conditions* (H1) *and* (H2) *and let Γ be defined according to* (4.6), (4.7). *The operator F defined by* (4.17) *satisfies the condition* (A3) *with $X = H^1(\mathbb{R})$ and $Y = H^{-1}(\mathbb{R})$. Furthermore, for $v \in X$, $D_sF(0, v) = D_sH(0, v) \in Y$ is the distribution defined by the following formulae for $w \in X$:*

$$\langle D_sF(0, v), w\rangle = \langle D_sH(0, v), w\rangle$$

$$= L(-\infty)\int_{-\infty}^{0} |x|^{-\beta}[|v|^{\sigma}vw](x)\,dx - L(+\infty)\int_{0}^{\infty} |x|^{-\beta}[|v|^{\sigma}vw](x)\,dx$$

in case (H2)(i),

$$= L(+\infty)\left[\int_{-1}^{1} \ln|x|\{|v|^{\sigma}vw\}'(x)\,dx - \int_{|x|\geq 1} x^{-1}[|v|^{\sigma}vw](x)\,dx\right]$$

$$+ \int_{-\infty}^{\infty} \psi(x)\,dx\,|v(0)|^{\sigma}v(0)w(0) \quad \textit{in case } (\text{H2})(\text{ii}),$$

$$= \frac{L(-\infty)}{(1-\beta)}\int_{-\infty}^{0} |x|^{1-\beta}[|v|^{\sigma}vw]'(x)\,dx + \frac{L(+\infty)}{(1-\beta)}\int_{0}^{\infty} |x|^{1-\beta}[|v|^{\sigma}vw]'(x)\,dx$$

in case (H3)(iii).

These formulae hold in the case where f has the form (4.1). *The corresponding results for the case* (4.1′) *are obtained by replacing $|v|^{\sigma}v$ by $|v|^{\sigma+1}$.*

Proof. We treat the case where f has the form (4.1). Since $\sigma > 1$, it follows by standard arguments that

$$G \text{ and } H(s, \cdot) \in C^1(X, Y) \quad \text{for all } s \in J = (-\Gamma(1), \Gamma(1)).$$

For the regularity of H with respect to s we deal in some detail with the case (H2)(i). For the other cases we simply mention the essential modifications that are necessary. To simplify the integral expressions it is convenient to set

$$L(x) = \mp L(\pm\infty) \quad \text{for } x \gtrless 0. \tag{4.18}$$

Suppose that R satisfies (H2)(i) and set

$$A = \max\{|L(\pm\infty)|\}, \qquad B = 2A/(1-\beta).$$

Then for $v, w \in X$,

$$\begin{aligned}\left|\int_{-\infty}^{\infty} L(x)|x|^{-\beta}[|v|^{\sigma}vw](x)\,dx\right| &\\ \leqq \| |v|^{\sigma}vw|_{\infty}\int_{-1}^{1}|L(x)|\,|x|^{-\beta}\,dx + A|v^{\sigma}|_{\infty}\int_{|x|\geqq 1}|v|\,|w|\,dx &\\ \leqq B|v|_{\infty}^{\sigma+1}|w|_{\infty} + A|v|_{\infty}^{\sigma}|v|_2|w|_2 &\\ \leqq (A+B)\|v\|_X^{\sigma+1}\|w\|_X. &\end{aligned} \tag{4.19}$$

Hence the form $\langle E_1(v), w\rangle = \int_{-\infty}^{\infty} L(x)|x|^{-\beta}[|v|^{\sigma}vw](x)\,dx$ defines an element $E_1(v) \in X^* = Y$ and $\|E_1(v)\|_Y \leqq (A+B)\|v\|_X^{\sigma+1}$. Similar estimates show that $E_1 \in C^1(X, Y)$ and

$$\langle DE_1(v)z, w\rangle = \int_{-\infty}^{\infty} L(x)|x|^{-\beta}(\sigma+1)|v|^{\sigma}zw\,dx \quad \text{for } v, z, w \in X.$$

Furthermore, for $s > 0$ and $v, w \in X$,

$$\begin{aligned}\left|\left\langle \frac{H(s,v)}{s} - E_1(v), w\right\rangle\right| &\leqq \|w\|_{\infty}\{A_1(s,v) + A_2(s,v)\} \\ &\leqq \|w\|_X\{A_1(s,v) + A_2(s,v)\}\end{aligned} \tag{4.20}$$

where

$$A_1(s,v) = \int_{|x|\leqq 2\gamma(s)} \left|R\left(\frac{x}{\gamma(s)}\right)\frac{|x|^{\beta}}{s} - L(x)\right| |x|^{-\beta}|v|^{\sigma+1}\,dx,$$

$$A_2(s,v) = \int_{|x|\geqq 2\gamma(s)} \left|R\left(\frac{x}{\gamma(s)}\right)\frac{|x|^{\beta}}{s} - L(x)\right| |x|^{-\beta}|v|^{\sigma+1}\,dx.$$

Clearly,

$$A_1(s,v) \leqq |v|_{\infty}^{\sigma+1}\left\{|R|_{\infty}\frac{2\gamma(s)}{s} + A\int_{|x|\leqq 2\gamma(s)}|x|^{-\beta}\,dx\right\}$$

and so $\lim_{s\to 0+} A_1(s,v) = 0$, since $\beta \in (0,1)$ and

$$\lim_{s\to 0+}\frac{\gamma(s)}{s} = \lim_{s\to 0+}\gamma'(s) = \lim_{s\to 0+}\gamma(s)^{1-\beta}[-\ln\gamma(s)]^{-\nu} = 0 \quad \text{by (4.13).}$$

For $|x| \geqq 2\gamma(s)$,

$$\begin{aligned}\left|R\left(\frac{x}{\gamma(s)}\right)\right|\frac{|x|^{\beta}}{s} = \frac{\gamma(s)^{\beta}}{s}\left|R\left(\frac{x}{\gamma(s)}\right)\right|\left|\frac{x}{\gamma(s)}\right|^{\beta}\left[\ln\left|\frac{x}{\gamma(s)}\right|\right]^{-\nu}\left[\ln\left|\frac{x}{\gamma(s)}\right|\right]^{\nu} &\\ \leqq \frac{\gamma(s)^{\beta}}{s}C\left[\ln\left|\frac{x}{\gamma(s)}\right|\right]^{\nu} \quad \text{by (4.8)} &\end{aligned}$$

where

$$\begin{aligned}C = \sup_{|z|\geqq 2} R(z)|z|^{\beta}[\ln|z|]^{-\nu} < \infty &\\ \leqq \frac{\gamma(s)^{\beta}}{s}C[-\ln\gamma(s)]^{\nu}[1+\ln|x|]^{\nu} &\end{aligned}$$

provided that $-\ln \gamma(s) \geqq 1$ since $\nu \geqq 0$. But by (4.13), $\lim_{s\to 0+} (\gamma(s)^\beta/s)[-\ln \gamma(s)]^\nu = \beta$ and $\lim_{s\to 0+} -\ln \gamma(s) = \infty$. Hence there exists $s_0 > 0$ such that

$$\left| R\left(\frac{x}{\gamma(s)}\right)\right| \frac{|x|^\beta}{s} \leqq C(\beta+1)[1+\ln|x|]^\nu \tag{4.21}$$

for $|x| \geqq 2\gamma(s)$ and $0 < s < s_0$.

Furthermore, for $x \neq 0$, by l'Hôpital's rule,

$$\begin{aligned}
&\lim_{s\to 0+} R\left(\frac{x}{\gamma(s)}\right)\frac{|x|^\beta}{s} \\
&= -\lim_{s\to 0+} R'\left(\frac{x}{\gamma(s)}\right)\frac{x|x|^\beta\gamma'(s)}{\gamma(s)^2} \\
&= -\frac{x}{|x|}\lim_{s\to 0+} R'\left(\frac{x}{\gamma(s)}\right)\left|\frac{x}{\gamma(s)}\right|^{\beta+1}\left[\ln\left|\frac{x}{\gamma(s)}\right|\right]^{-\nu}\left[\ln\left|\frac{x}{\gamma(s)}\right|\right]^{\nu}\gamma'(s)\gamma(s)^{\beta-1} \\
&= \mp L(\pm\infty) \quad \text{for } x \gtrless 0 \quad \text{by (4.3)}
\end{aligned} \tag{4.22}$$

since by (4.13)

$$\lim_{s\to 0+}\left[\ln\left|\frac{x}{\gamma(s)}\right|\right]^\nu \gamma'(s)\gamma(s)^{\beta-1} = \lim_{s\to 0+}\left[\frac{\ln|x| - \ln\gamma(s)}{-\ln\gamma(s)}\right]^\nu = 1.$$

Now, denoting the characteristic function of $\{x: |x| \geqq 2\gamma(s)\}$ by χ_s, we have that

$$A_2(s, v) = \int_{-\infty}^{\infty} \chi_s(x)\left| R\left(\frac{x}{\gamma(s)}\right)\frac{|x|^\beta}{2} - L(x)\right| |x|^{-\beta}|v(x)|^{\sigma+1}\, dx$$

where the integrand is bounded by

$$\{C(\beta+1)[1+\ln|x|]^\nu + A\}|x|^{-\beta}|v(x)|^{\sigma+1}$$

provided that $0 < s < s_0$. Hence the Dominated Convergence Theorem applies and we conclude from (4.22) that $\lim_{s\to 0+} A_2(s, v) = 0$. Returning to (4.20), we have that

$$\lim_{s\to 0}\left\| \frac{H(s, v)}{s} - E_1(v)\right\|_Y = 0,$$

proving that H is differentiable with respect to s at $s = 0$ and that $D_sH(0, v) = E_1(v)$. For $s > 0$,

$$\langle D_sH(s, v), w\rangle = -\int_{-\infty}^{\infty} R'\left(\frac{x}{\gamma(s)}\right)\frac{x\gamma'(s)}{\gamma(s)^2}[|v|^\sigma vw](x)\, dx$$

and by arguments similar to those above we find that

$$\lim_{s\to 0+} \|D_sH(s, v) - E_1(v)\|_Y = 0.$$

From this and the continuity of $E_1: X \to Y$, it follows that $H \in C^1(J \times X, Y)$.

The discussion of the derivatives D_sD_vH and D_vD_sH is similar to that for D_sH, and we find that

$$D_sD_vH(0, v) = D_vD_sH(0, v) = DE_1(v) \quad \forall v \in X.$$

In this way we see that F satisfies (A1) when (H2)(i) holds.

Turning to the case where r satisfies (H2)(ii), we note that it is enough to establish the result for the following special cases:

$$R(x) = x/(1+x^2), \tag{4.23}$$

$$R(x) = \psi(x). \tag{4.24}$$

We discuss (4.23) first. Defining a bilinear form by

$$\langle E_2(v), w\rangle = -\int_{-1}^{1} \ln|x|\{|v|^\sigma vw\}'(x)\,dx + \int_{|x|\geqq 1} x^{-1}\{|v|^\sigma vw\}(x)\,dx \tag{4.25}$$

we show that for each $v \in X$, it defines a bounded linear functional on X. In fact,

$$\begin{aligned}\left|\int_{-1}^{1} \ln|x|\{|v|^\sigma vw\}'(x)\,dx\right| &\leqq \int_{-1}^{1} |\ln|x||\{(\sigma+1)|v|^\sigma|v'||w| + |v|^{\sigma+1}|w'|\}\,dx \\ &\leqq (\sigma+1)|v|_\infty^\sigma |w|_\infty D|v'|_2 + |v|_\infty^{\sigma+1} D|w'|_2 \\ &\leqq D(\sigma+2)\|v\|_X^{\sigma+1}\|w\|_X\end{aligned}$$

where

$$D = \left\{\int_{-1}^{1} (\ln|x|)^2\,dx\right\}^{1/2}$$

and

$$\left|\int_{|x|\geqq 1} x^{-1}\{|v|^\sigma vw\}(x)\,dx\right| \leqq |v|_\infty^\sigma |v|_2 |w|_2 \leqq \|v\|_X^{\sigma+1}\|w\|_X.$$

Thus $E_2(v) \in X^* = Y$ for all $v \in X$ and by similar arguments $E_2 \in C^1(X, Y)$. For $s > 0$ and $v, w \in X$,

$$\begin{aligned}\left\langle \frac{H(s,v)}{s}, w\right\rangle &= \int_{-\infty}^{\infty} \frac{1}{s} R\left(\frac{x}{s}\right)[|v|^\sigma vw](x)\,dx \\ &= \int_{-1}^{1} \frac{x}{s^2+x^2}\{|v|^\sigma vw(x) - |v|^\sigma vw(0)\}\,dx \\ &\quad + \int_{|x|\geqq 1} \frac{x^2}{s^2+x^2} x^{-1}\{|v|^\sigma vw\}(x)\,dx \\ &= -\int_{-1}^{1} \frac{1}{2}\ln(s^2+x^2)\{|v|^\sigma vw\}'(x)\,dx \\ &\quad + \frac{1}{2}\ln(s^2+1)\{|v|^\sigma vw(1) - |v|^\sigma vw(-1)\} \\ &\quad + \int_{|x|\geqq 1} \frac{x^2}{s^2+x^2} x^{-1}\{|v|^\sigma vw\}(x)\,dx.\end{aligned}$$

Using this expression, it is now easy to show that

$$\lim_{s\to 0+} \left\| \frac{H(s,v)}{s} - E_2(v)\right\|_Y = 0$$

and so $D_s H(0, v) = E_2(v)$ for all $v \in X$.

The other properties are established in a similar fashion for (4.23) and so we now pass to the discussion of (4.24).

In this case we set

$$\langle E_3(v), w\rangle = \int_{-\infty}^{\infty} \psi(x)\,dx|v(0)|^\sigma v(0)w(0).$$

Since

$$|\langle E_3(v), w\rangle| \leqq \left| \int_{-\infty}^{\infty} \psi(x)\, dx \right| |v|_\infty^{\sigma+1} |w|_\infty$$
$$\leqq \left| \int_{-\infty}^{\infty} \psi(x)\, dx \right| \|v\|_X^{\sigma+1} \|w\|_X$$

we see that $E_3(v) \in Y$ for all $v \in X$ and by standard arguments $E_3 \in C^1(X, Y)$.

Now for $s > 0$ and $v, w \in X$,

$$\left\langle \frac{H(s, v)}{s}, w \right\rangle = \int_{-\infty}^{\infty} \psi\left(\frac{x}{s}\right) \frac{1}{s} [|v|^\sigma vw](x)\, dx$$
$$= -\int_{-\infty}^{\infty} P\left(\frac{x}{s}\right) [|v|^\sigma vw]'(x)\, dx \quad \text{where } P(x) = \int_{-\infty}^{x} R(t)\, dt$$

and so

$$\left\langle \frac{H(s, v)}{s} - E_3(v), w \right\rangle = -\int_{-\infty}^{\infty} \left[P\left(\frac{x}{s}\right) - N(x) \right] [|v|^\sigma vw]'(x)\, dx$$

where

$$N(x) = \begin{cases} 0 & \text{if } x < 0, \\ \displaystyle\int_{-\infty}^{\infty} \psi(t)\, dt & \text{if } x > 0. \end{cases}$$

Since $\lim_{s \to 0+} P(x/s) - N(x) = 0$ for all $x \neq 0$ it is now easy to show, using dominated convergence, that

$$\lim_{s \to 0+} \left\| \frac{H(s, v)}{s} - E_3(v) \right\|_Y = 0.$$

This proves that $D_s H(0, v) = E_3(v)$ when (4.24) holds and the remaining properties of H are established in a similar fashion.

Finally, we consider the case where r satisfies (H2)(iii). As usual we begin by studying the bilinear form defined by

$$\langle E_4(v), w\rangle = \frac{L(-\infty)}{(1-\beta)} \int_{-\infty}^{0} |x|^{1-\beta} [|v|^\sigma vw]'(x)\, dx$$
$$+ \frac{L(+\infty)}{(1-\beta)} \int_{0}^{\infty} |x|^{1-\beta} [|v|^\sigma vw]'(x)\, dx. \tag{4.25}$$

Now setting $\left| \int_{-\infty}^{0} |x|^{1-\beta} [|v|^\sigma vw]'(x)\, dx \right| = I_1$, we have that

$$I_1 \leqq \int_{-\infty}^{0} |x|^{1-\beta} \{(\sigma+1)|v|^\sigma |v'||w| + |v|^{\sigma+1}|w'|\}(x)\, dx$$
$$\leqq \int_{-\infty}^{-1} \{(\sigma+1)|v|^\sigma |v'||w| + |v|^{\sigma+1}|w'|\}(x)\, dx$$
$$+ \int_{-1}^{0} |x|^{1-\beta} \{(\sigma+1)|v|^\sigma |v'||w| + |v|^{\sigma+1}|w'|\}(x)\, dx$$
$$\leqq (\sigma+1)|v|_\infty^\sigma |v'|_2 |w|_2 + |v|_\infty^\sigma |v|_2 |w'|_2$$
$$+ (\sigma+1)|v|_\infty^\sigma |w|_\infty F |v'|_2 + |v|_\infty^{\sigma+1} F |w'|_2$$

where

$$F=\left\{\int_{-1}^{0}|x|^{2(1-\beta)}\,dx\right\}^{1/2},$$

$$I_1\leqq(\sigma+2)(F+1)\|v\|_X^{\sigma+1}\|w\|_X.$$

The second integral in (4.25) is estimated in the same way, so we can conclude that $E_4\in C^1(X, Y)$. For $s>0$ and $v, w\in X$, we have that

$$\left\langle\frac{H(s,v)}{s}, w\right\rangle=\int_{-\infty}^{\infty}R\left(\frac{x}{\gamma(s)}\right)\frac{1}{s}[|v|^{\sigma}vw](x)\,dx$$

$$=-\int_{-\infty}^{\infty}P\left(\frac{x}{\gamma(s)}\right)\frac{\gamma(s)}{s}[|v|^{\sigma}vw]'(x)\,dx$$

where $P(x)=\int_{-\infty}^{x}R(t)\,dt$. Since $\int_{-\infty}^{\infty}R(t)\,dt=0$ we can use l'Hôpital's rule and obtain

$$\lim_{s\to0+}P\left(\frac{x}{\gamma(s)}\right)\frac{\gamma(s)}{s}=\lim_{s\to0+}\left\{-R\left(\frac{x}{\gamma(s)}\right)\frac{x\gamma'(s)}{\gamma(s)}+P\left(\frac{x}{\gamma(s)}\right)\gamma'(s)\right\}$$

$$=-\lim_{s\to0+}\left\{\frac{x}{\gamma(s)}R\left(\frac{x}{\gamma(s)}\right)-P\left(\frac{x}{\gamma(s)}\right)\right\}\gamma'(s)=-\frac{L(\pm\infty)}{1-\beta}|x|^{\beta-1}$$

by (4.12) and (4.13).

Arguing much as in (H2)(i), the Dominated Convergence Theorem can now be used to show that

$$\lim_{s\to0+}\left\|\frac{H(s,v)}{s}-E_4(v)\right\|_Y=0\quad\forall v\in X,$$

and the proof is completed in the usual way.

LEMMA 4.2. *Let f satisfy conditions* (H1) *and* (H2) *and consider the function $F\colon J\times X\to Y$ defined by* (4.17) *via the transformation* (4.6), (4.7). *Then*

$$\{v\in X\backslash\{0\}\colon F(0,v)=0\}=\{\pm T_\tau v_0\colon\tau\in\mathbb{R}\}\quad\textit{in case (4.1)}$$

$$=\{T_\tau v_0\colon\tau\in\mathbb{R}\}\quad\textit{in case (4.1')}$$

where

$$v_0(x)=\left\{\frac{\sigma+2}{2Q}\right\}^{1/\sigma}ch^{-2/\sigma}\left(\frac{\sigma x}{2}\right)\quad\textit{for } x\in\mathbb{R}. \tag{4.26}$$

In particular, (A4) *is satisfied by v_0, which is even and positive.*

Proof. If $v\in X$ and $F(0, v)=0$, it follows from Lemma 2.1 that $v\in H^3(\mathbb{R})\cap C^3(\mathbb{R})$ and v satisfies (2.19). By (4.2), in the case (4.1), (2.19) is

$$v''(x)-v(x)+Q|v(x)|^{\sigma}v(x)=0\quad\forall x\in\mathbb{R}. \tag{4.27}$$

By direct calculation we show that

$$\left\{v\in C^2(\mathbb{R})\backslash\{0\}\colon\text{(4.27) is satisfied and }\lim_{|x|\to\infty}v(x)=0\right\}=\{\pm T_\tau v_0\colon\tau\in\mathbb{R}\}.$$

Since $v_0, v_0'\in X$ this completes the proof in case (4.1). The case (4.1') is similar.

By Lemmas 4.1 and 4.2, we know that when r satisfies (H1) and (H2) and Γ is defined by (4.6), (4.7), the conditions (A1)-(A4) of § 3 are all fulfilled, and it is sufficient

to consider the Melnikov function generated by the function v_0 given in (4.26). From Lemma 4.1 and definition (3.13), the Melnikov function M is given by

$$M(\tau) = \langle D_s H(0, T_\tau v_0), T_\tau v_0' \rangle$$

since $i(x) \equiv 1$ by (4.2).

When we set

$$m_\beta(\tau) = \begin{cases} \displaystyle\int_0^\infty \frac{x^{1-\beta}}{(1-\beta)} [v_0^{\sigma+1} v_0']'(x+\tau)\, dx & \text{for } \beta \neq 1, \\ \displaystyle\int_0^\infty \ln x [v_0^{\sigma+1} v_0']'(x+\tau)\, dx & \text{for } \beta = 1, \end{cases} \tag{4.28}$$

where v_0 is given by (4.26), it follows from Lemma 4.1 that the Melnikov function for (4.1) or (4.1′) is given by

$$M(\tau) = L(-\infty) m_\beta(-\tau) + L(+\infty) m_\beta(\tau) \quad \text{in cases (H1)(i) and (H1)(iii),} \tag{4.29}$$

$$= L(+\infty)[m_1(-\tau) + m_1(\tau)] + \int_{-\infty}^{\infty} \psi(x)\, dx\, v_0(\tau)^{\sigma+1} v_0'(\tau) \tag{4.30}$$

in case (H2)(ii).

In connection with (4.28), it is worth noting that

$$[v_0^{\sigma+1} v_0']' = \frac{(v_0^{\sigma+2})''}{\sigma+2} \tag{4.31}$$

$$= (\sigma+2) v_0^{\sigma+2} \left\{ 1 - \frac{(3\sigma+4)}{(\sigma+2)^2} Q v_0^\sigma \right\}, \tag{4.32}$$

and furthermore $m_\beta(\tau) = dn_\beta(\tau)/d\tau$, where

$$n_\beta(\tau) = \begin{cases} \displaystyle\int_0^\infty \frac{x^{1-\beta}}{(1-\beta)} [v_0^{\sigma+1} v_0'](x+\tau)\, dx & \text{for } \beta \neq 0, \\ \displaystyle\int_0^\infty \ln x [v_0^{\sigma+1} v_0'](x+\tau)\, dx & \text{for } \beta = 1. \end{cases}$$

It is easy to prove that $\lim_{|\tau| \to \infty} n_\beta(\tau) = 0$ for $0 < \beta < 3/2$ and

$$M(\tau) = \frac{d}{d\tau} \{ -L(-\infty) n_\beta(-\tau) + L(+\infty) n_\beta(\tau) \} \quad \text{in cases (H2)(i) and (H2)(iii)}$$

$$= \frac{d}{d\tau} \left\{ -L(+\infty) n_1(-\tau) + L(+\infty) n_1(\tau) + \int_{-\infty}^{\infty} \psi(x)\, dx \frac{v_0(\tau)^{\sigma+2}}{(\sigma+2)} \right\}$$

in case (H2)(ii).

Thus in all cases M is the derivative of a function converging to zero as $\tau \to \pm\infty$.

This ensures that in all cases there exists at least one $\tau_0 \in \mathbb{R}$ such that $M(\tau_0) = 0$. By Theorem 3.2 we must seek the simple zeros of M.

In terms of bifurcation in the sense of (1.5) for the problem (1.1), (1.2) we obtain the following result.

THEOREM 4.3. *Let f be of the form* (4.1) *where* (H1) *and* (H2) *are satisfied. Suppose that*

$$M(\tau_0) = 0 \quad \textit{and} \quad M'(\tau_0) \neq 0 \tag{4.33}$$

where M is the Melnikov function defined by (4.29), (4.30). *Then there is bifurcation in $L^p(\mathbb{R})$ for* (1.1), (1.2) *in the sense of* (1.5) *provided that $p \geqq \min\{2, \sigma/2\}$. Furthermore the branch of solutions has the form given by Theorem* 3.2 *with $\alpha = 2/\sigma$, where v_0 is defined by* (4.26) *and Γ by* (4.6), (4.7).

This result is an immediate consequence of Theorem 3.2 and Lemmas 4.1 and 4.2. In general it seems that the location of simple zeros of M must be treated by numerical integration. See the Appendix. However, there are some cases in which the solution is easy. For example, if

$$R \in L^1(\mathbb{R}) \quad \text{with} \lim_{|x|\to\infty} xR(x) = 0 \quad \text{and} \quad \int_{-\infty}^{\infty} R(x)\,dx \neq 0, \tag{4.34}$$

then we are in case (H2)(ii) with $L(+\infty) = 0$ and

$$M(\tau) = \int_{-\infty}^{\infty} R(x)\,dx\, v_0(\tau)^{\sigma+1} v_0'(\tau).$$

Hence,

$$M(0) = 0 \quad \text{and} \quad M'(0) = -\int_{-\infty}^{\infty} R(x)\,dx \frac{\sigma}{2}\left[\frac{\sigma+2}{Q}\right]^{(\sigma+2)/\sigma} \neq 0 \quad \text{by (4.32).} \tag{4.35}$$

Another easy case is that in which R satisfies (H2)(i) with

$$L(+\infty) = -L(-\infty) \neq 0. \tag{4.36}$$

In this case,

$$M(0) = L(+\infty)[-m_\beta(0) + m_\beta(0)] = 0,$$

and

$$\begin{aligned} M'(0) = 2L(+\infty)m_\beta'(0) &= L(+\infty)\int_0^\infty \frac{x^{1-\beta}}{(1-\beta)}\left\{\frac{v_0^{\sigma+2}}{\sigma+2}\right\}'''(x)\,dx \\ &= -L(+\infty)\int_0^\infty x^{-\beta}\left\{\frac{v_0^{\sigma+1}}{\sigma+2}\right\}''\,dx \\ &= -\frac{L(+\infty)}{(\sigma+2)}\lim_{\varepsilon\to 0+}\int_\varepsilon^\infty x^{-\beta}\{v_0^{\sigma+2}\}''\,dx \\ &= -\frac{L(+\infty)}{(\sigma+2)}\lim_{\varepsilon\to 0+}\left\{x^{-\beta}(v_0^{\sigma+2}(x))'\Big|_\varepsilon^\infty + \beta\int_\varepsilon^\infty x^{-\beta-1}\{v_0^{\sigma+2}\}'\,dx\right\} \\ &= -\frac{\beta L(+\infty)}{(\sigma+2)}\int_0^\infty x^{-\beta-1}\{v_0(x)^{\sigma+1}\}'\,dx \quad \text{since } v_0'(0) = 0 \\ &\neq 0 \quad \text{since } (v_0(x)^{\sigma+1})' < 0 \quad \text{for } x > 0. \end{aligned} \tag{4.37}$$

We end this section with two typical examples showing how the various cases in (H2) arise.

Example 4.4. Let f be of the form (4.1) with

$$r(x) = Q + C(1+x^2)^{-\gamma} \quad \text{for } x \in \mathbb{R}$$

where $C \in \mathbb{R}\backslash\{0\}$, $Q > 0$, and $\gamma > 0$.

First we note that since r is even, this case is covered by the result in [14]. In terms of the present discussion we note that (H1) is satisfied provided that $\sigma > 1$. Then

(H2)(i) holds for $\beta \leqq \min\{2\gamma, 1\}$, where $L(\pm\infty) = 0$ unless $\beta = 2\gamma < 1$ and $\nu = 0$, in which case $L(\pm\infty) = \pm 2\gamma C$. Hence for $0 < \gamma < \frac{1}{2}$, we use $\beta = 2\gamma$ and $\nu = 0$ in (H2)(i) to obtain the Melnikov function

$$M(\tau) = -4\gamma C \int_0^\infty x^{-\beta}[v_0^{\sigma+1} v_0'](x+\tau)\, dx$$

and by (4.36), $M(0) = 0$ and $M'(0) \neq 0$. If $\gamma > \frac{1}{2}$, then (H2)(ii) is satisfied with $L(+\infty) = 0$ and we obtain the Melnikov function

$$M(\tau) = C \int_{-\infty}^\infty (1+x^2)^{-\gamma}\, dx\, v_0(\tau)^{\sigma+1} v_0'(\tau).$$

By (4.34), $M(0) = 0$ and $M'(0) \neq 0$. None of the results in this paper covers the case $\gamma = \frac{1}{2}$, but due to the evenness of R the result in [14] applies.

Example 4.5. Let f be of the form (4.1) with

$$r(x) = Q + Cx(1+x^2)^{-\gamma} \quad \text{for } x \in \mathbb{R}$$

where $C \in \mathbb{R}\backslash\{0\}$, $Q > 0$, and $\gamma > \frac{1}{2}$.

In this case the results in [14] do not apply since r is not even. In terms of the present discussion (H1) holds provided that $\sigma > 1$, and (H2)(i) is satisfied provided that $\beta \leqq \min\{2\gamma - 1, 1\}$, where $L(\pm\infty) = 0$ unless $\beta = 2\gamma - 1 < 1$ and $\nu = 0$ in which case $L(\pm\infty) = (1-2\gamma)C$. Thus we get a nontrivial Melnikov function by using $\beta = 2\gamma - 1$ and $\nu = 0$ in (H2)(i) provided that $\frac{1}{2} < \gamma < 1$. For $\gamma = 1$ we apply the case (H2)(ii) with

$$L(+\infty) = -1 \quad \text{and} \quad \psi \equiv 0.$$

For $\gamma > 1$ the case (H2)(ii) again applies with $L(+\infty) = 0$, but we obtain a trivial Melnikov function since $\int_{-\infty}^\infty R(x)\, dx = 0$. However, for $\gamma > 1$ we see that r satisfies (H2)(iii) provided that

$$\beta \leqq \min\{2\gamma - 1, \tfrac{3}{2}\}$$

where $L(\pm\infty) = 0$ unless $\beta = 2\gamma - 1 < 3/2$ and $\nu = 0$, in which case $L(\pm\infty) = (1-2\gamma)C$. Thus for $1 < \gamma < 5/4$ we obtain a nontrivial Melnikov function by using (H2)(iii) with $\beta = 2\gamma - 1$ and $\nu = 0$.

5. More general nonlinearities. In this section we discuss nonlinearities f in (1.1) that are sums of terms of the form (4.1), and we also allow perturbations containing a factor u'. The basic idea is to select the dominant contribution, together with the appropriate change of variables, ensuring that it leads to a nontrivial Melnikov function as in § 4. Then we show that the remaining terms are regular (in the sense of (A3)) and lead to Melnikov functions that are identically zero for this same change of variables. In the simplest cases the dominant term is the one with the smallest value of the exponent σ, so we discuss this kind of situation first. However, there are problems where the dominant term is not given by the smallest power of u, and we conclude with some examples of this kind.

We suppose that the nonlinearity f in (1.1) can be written as

$$f(\lambda, x, p, q) = f_0(x, p) + \sum_{i=1}^N \lambda^{n_i} f_i(x, p) q^{\delta_i} \tag{5.1}$$

where f_i is of the form (4.1) or (4.1′) for $0 \leqq i \leqq N$ with coefficient r_i and exponent σ_i. Setting $r_0 = r$ and $\sigma_0 = \sigma$, we suppose that f_0 satisfies (H1) and (H2) with exponent β in (H2) ($\beta = 1$ for the case (H2)(ii)) and $R(x) = Q - r(x)$.

For $1 \leqq i \leqq N$ we suppose that

$$n_i \in \{0, 1, 2, \cdots\} = \mathbb{N}, \qquad \delta_i \in \{0, 1\}$$

and we set

$$\mu_i = 2n_i - 2 + \frac{2\sigma_i}{\sigma} + \delta_i\left(1 + \frac{2}{\sigma}\right).$$

For $\alpha = 2/\sigma$ we find that

$$k^{-(2+\alpha)} f\left(-k^2, \frac{x}{k}, k^{\alpha} p, k^{\alpha+1} q\right) = Q|p|^{\sigma} p + R\left(\frac{x}{k}\right)|p|^{\sigma} p + \sum_{i=1}^{N} (-1)^{n_i} k^{\mu_i} f_i\left(\frac{x}{k}, p\right) q^{\delta_i}$$

when f_0 satisfies (4.1), whereas if f_0 satisfies (4.1′) we simply replace $|p|^{\sigma} p$ by $|p|^{\sigma+1}$.

Let $\gamma = \Gamma^{-1}$ be the change of variables associated with r via (H2) and (4.6), (4.7). Let F_0: $J \times X \to Y$ be the operator associated with f_0 as in (4.17). Thus if f_0 satisfies (4.1)

$$F_0(s, v)(x) = v''(x) - v(x) + Q|v(x)|^{\sigma} v(x) + R\left(\frac{x}{\gamma(s)}\right)|v(x)|^{\sigma} v(x)$$

and if f_0 satisfies (4.1′) we simply replace $|v|^{\sigma} v$ by $|v|^{\sigma+1}$. As usual $X = H^1(\mathbb{R})$ and $Y = H^{-1}(\mathbb{R})$. By Lemmas 4.1 and 4.2, F satisfies (A3).

For $1 \leqq i \leqq N$, we set

$$K_i(k, v)(x) = (-1)^{n_i} f_i\left(\frac{x}{k}, v(x)\right) v'(x)^{\delta_i}$$

and

$$H_i(s, v) = \begin{cases} \gamma(s)^{\mu_i} K_i(\gamma(s), v) & \text{for } 0 < s < \Gamma(1), \\ 0 & \text{for } s = 0, \\ -H_i(-s, v) & \text{for } -\Gamma(1) < s < 0. \end{cases}$$

The object now is to give conditions implying that H_i satisfies (A3) and $D_s H_i(0, v) \equiv 0$ for $1 \leqq i \leqq N$. From this it will follow that f satisfies the conditions (A1)–(A4) with v_0 given by (4.26) and Melnikov function given by (4.29), (4.30). This objective is attained by imposing the following conditions.

(H3) For $1 \leqq i \leqq N$, $r_i \in C^1(\mathbb{R}) \cap L^{\infty}(\mathbb{R})$ with $|x r_i'(x)| \leqq C$ for all $x \in \mathbb{R}$ with $\mu_i > \beta$ unless $r_i \equiv 0$. Furthermore, $f_1(x, p) = r_1(x) p$ and $\delta_1 = 1$, whereas for $2 \leqq i \leqq N$, f_i has the form (4.1) or (4.1′) with $\sigma_i > 1$ or $\sigma_i \geqq 1$ and $\delta_i \in \{0, 1\}$.

Under the above hypotheses it is easy to see that f, given by (5.1), satisfies (A1) and (A2). Also for $s > 0$,

$$D_s H_i(s, v)(x) = (-1)^{n_i} \gamma(s)^{\mu_i - 1} \gamma'(s) \cdot \left\{\mu_i r_i\left(\frac{x}{\gamma(s)}\right) - \frac{x}{\gamma(s)} r_i'\left(\frac{x}{\gamma(s)}\right)\right\} |v(x)|^{\sigma_i} v(x) v'(x)^{\delta_i}$$

if f_i has the form (4.1), and in the case (4.1′) we simply replace $|v|^{\sigma_i} v$ by $|v|^{\sigma_i + 1}$. However, $\mu_i > \beta > 0$, and by (4.13)

$$\lim_{s \to 0+} \gamma(s)^{\mu_i - 1} \gamma'(s) = \lim_{s \to 0+} \gamma(s)^{\mu_i - \beta} [-\ln \gamma(s)]^{-\nu} = 0.$$

Since r_i and $x r_i'(x) \in L^{\infty}(\mathbb{R})$ it follows easily that $\|D_s H(s, v)\|_Y \to 0$ as $s \to 0$ and by a similar argument that

$$\left\| \frac{H(s, v)}{s} \right\|_Y \to 0 \quad \text{as } s \to 0.$$

Thus we see that (H3) implies that $H_i \in C^1(J \times X, Y)$ with $D_s H_i(0, v) \equiv 0$ for $1 \leqq i \leqq N$. The information concerning second partial derivatives for H_i required by (A3) is checked by the same argument. Hence we have established the following result.

THEOREM 5.1. *Let f be of the form* (5.1) *and suppose that* (H1), (H2) *are satisfied by r and σ and that* (H3) *holds. Let $\gamma = \Gamma^{-1}$ be the change of variables defined by* (4.6), (4.7) *via* (H2) *using r and σ. Then conditions* (A1)-(A4) *are fulfilled by f, and the Melnikov function M for f is determined by r and σ through* (4.26) *and* (4.29), (4.30). *In particular, there is bifurcation in $L^p(\mathbb{R})$ for* (1.1), (1.2) *in the sense of* (1.5) *provided that* (4.33) *holds and $p \geqq \min\{2, \sigma/2\}$.*

When $n_i = 0$ for $1 \leqq i \leqq N$, $\delta_i = 0$ for $2 \leqq i \leqq N$, and $r_1 \equiv 0$, it follows from (H3) that $\sigma_i > \sigma$ for all $i \geqq 2$ and hence the dominant contribution to f comes from the smallest power σ. We end with some examples of situations where the dominant contribution comes from a higher power of u.

Example 5.2.

$$f(\lambda, x, p, q) = \frac{x^2}{1+x^2} p^3 + \frac{x}{(1+x^2)^2} p^2.$$

Setting

$$Q = 1, \quad R(x) = -\frac{1}{1+x^2}, \quad r_1(x) = \frac{x}{(1+x^2)^2},$$

we have $f(\lambda, x, p, q) = p^3 + R(x)p^3 + r_1(x)p^2$, where $\lim_{|x|\to\infty} R(x) = \lim_{|x|\to\infty} r_1(x) = 0$. Clearly, (A1) is satisfied and setting $\alpha = 1$ and $g(p, q) = p^2$, we find that (A2) is also satisfied. Furthermore, R and $r_1 \in L^1(\mathbb{R})$ with

$$\lim_{|x|\to\infty} xR(x) = \lim_{|x|\to\infty} xr_1(x) = 0.$$

Hence both R and r_1 satisfy (H2)(ii) with $L(\pm\infty) = 0$. Thus by Lemmas 4.1 and 4.2, (A3) is satisfied (with change of variable $\gamma(s) = s$) and

$$\langle D_s H(0, v), w\rangle = \int_{-\infty}^{\infty} R(x)\, dx\, v(0)^3 w(0) + \int_{-\infty}^{\infty} r_1(x)\, dx\, v(0)^2 w(0)$$
$$= -\pi v(0)^3 w(0) \quad \text{for all } w \in X.$$

Since $g(p, q) = p^3$ we see that (A4) holds with $v_0(x) = \sqrt{2}\, ch^{-1}x$ and the Melnikov function in this case is $M(\tau) = 4\pi sh\pi ch^{-5}\tau$. By (4.35), $M(0) = 0$ and $M'(0) \neq 0$.

Example 5.3.

$$f(\lambda, x, p, q) = \{1 + (1+x^2)^{-1/4}\}p^3 + (1+x^2)^{-1/2}p^2.$$

Clearly (A1) holds. Setting $\alpha = 1$ and $g(p, q) = p^3$, we verify easily that (A2) is satisfied. Setting $R(x) = (1+x^2)^{-1/4}$ and $r_1(x) = (1+x^2)^{-1/2}$, we find that

$$\lim_{x\to\pm\infty} R'(x)|x|^{3/2} = \mp\tfrac{1}{2} \quad \text{and} \quad \lim_{x\to\pm\infty} r_1'(x)|x|^{3/2} = 0.$$

Hence both R and r_1 satisfy (H2)(i) with $\beta = \frac{1}{2}$, $\nu = 0$, and

$$L(\pm\infty) = \mp\tfrac{1}{2} \quad \text{for } R,$$
$$L(\pm\infty) = 0 \quad \text{for } r_1.$$

It follows that (A3) is satisfied (with $\gamma(s) = s^2/4$) and that

$$\langle D_s H(0, v), w\rangle = \frac{1}{2}\int_{-\infty}^{\infty} |x|^{-1/2} v(x)^3 w(x)\, dx.$$

Again we are in a situation where (A4) holds with

$$v_0(x) = \sqrt{2}\, ch^{-1} x$$

and in this case the Melnikov function is

$$M(\tau) = -2 \int_{-\infty}^{\infty} |x|^{-1/2} \frac{sh(x+\tau)}{[ch(x+\tau)]^5}\, dx.$$

By (4.37), $M(0) = 0$ and $M'(0) < 0$.

Appendix. We present the graphs of some of the Melnikov functions encountered in § 4 (Figs. 1-6). They were obtained through numerical integration by François Meynard.

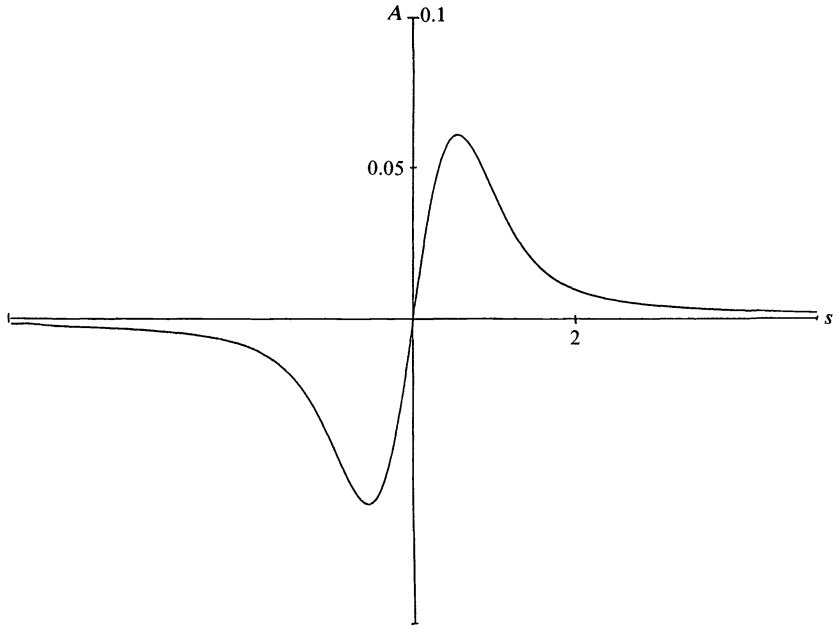


FIG. 1. $\sigma = 2$; $\beta = 0.5$; $L(-\infty) = -1$; $L(+\infty) = 1$. *One value for* s_0: $s_0 = 0$. *This agrees with* (4.36).

It is convenient to begin with some simplifications of the original formula. From (4.26), (4.28), and (4.32), we find that for $\beta \neq 1$,

$$m_\beta(\tau) = \int_0^{\infty} \frac{x^{1-\beta}}{(1-\beta)}\, w(x+\tau)\, dx$$

where

$$w(x) = \frac{(\sigma+2)^{2+2/\sigma}}{(2Q)^{1+2/\sigma}} \frac{1}{ch^{\alpha}(\sigma x/2)} \left\{ 1 - \frac{(1+1/\alpha)}{ch^2(\sigma x/2)} \right\}$$

and

$$\alpha = 2 + \frac{4}{\sigma}.$$

Hence there exists $C(Q, \sigma) > 0$ such that for $\beta \neq 1$,

$$M(\tau) = \frac{C(Q,\sigma)}{(1-\beta)} A\left(\frac{\sigma\tau}{2}\right)$$

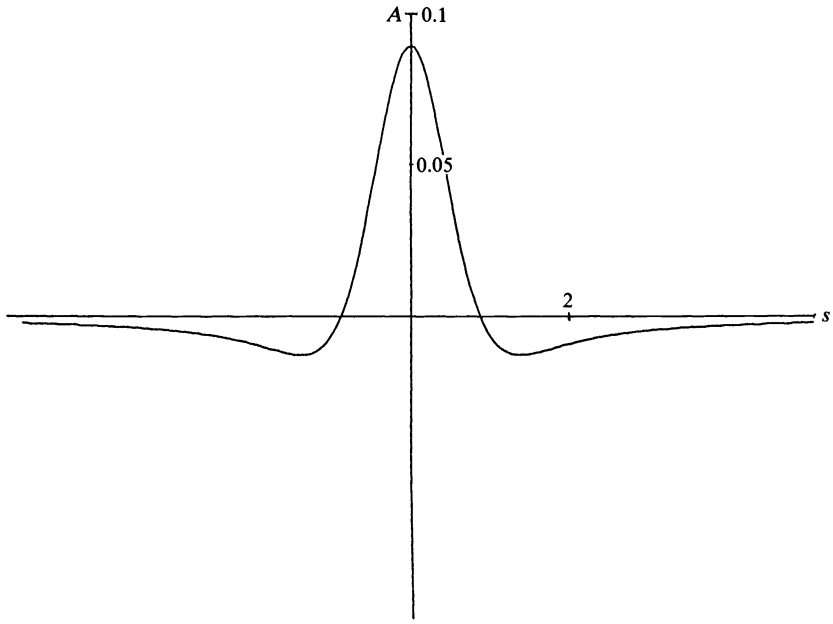


FIG. 2. $\sigma = 2$; $\beta = 0.5$; $L(-\infty) = 1$; $L(+\infty) = 1$. *Two values for* s_0: $s_0 \simeq +0.87$; $s_0 \simeq -0.87$.

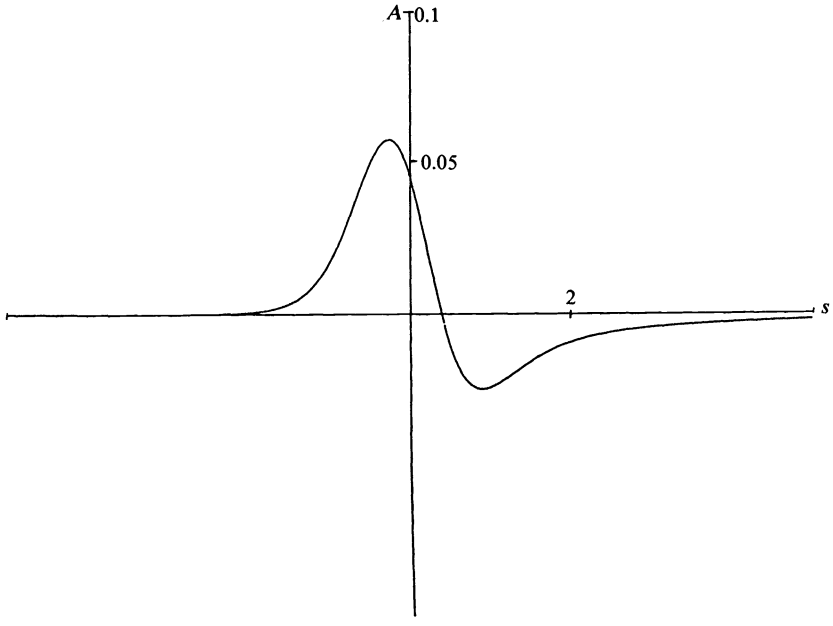


FIG. 3. $\sigma = 2$; $\beta = 0.5$; $L(-\infty) = 1$; $L(+\infty) = 0$. *One value for* s_0: $s_0 \simeq +0.37$.

where $A(s) = L(-\infty)B(-s) + L(+\infty)B(s)$ with

$$B(s) = \int_0^\infty x^{1-\beta} \frac{1}{ch^\alpha(x+s)} \left\{ 1 - \frac{(1+1/\alpha)}{ch^2(x+s)} \right\} dx.$$

Then (4.33) is equivalent to $\tau_0 = 2s_0/\sigma$, where $A(s_0) = 0$ and $A'(s_0) \neq 0$.

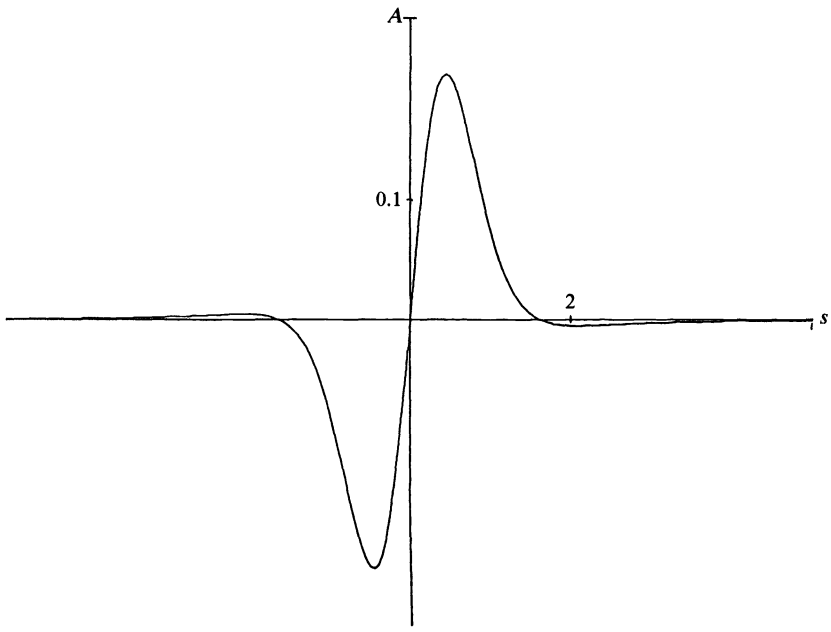


FIG. 4. $\sigma = 2$; $\beta = 1.25$; $L(-\infty) = -1$; $L(+\infty) = 1$. *Three values for* s_0: $s_0 = 0$; $s_0 \simeq +1.63$; $s_0 \simeq -1.63$.

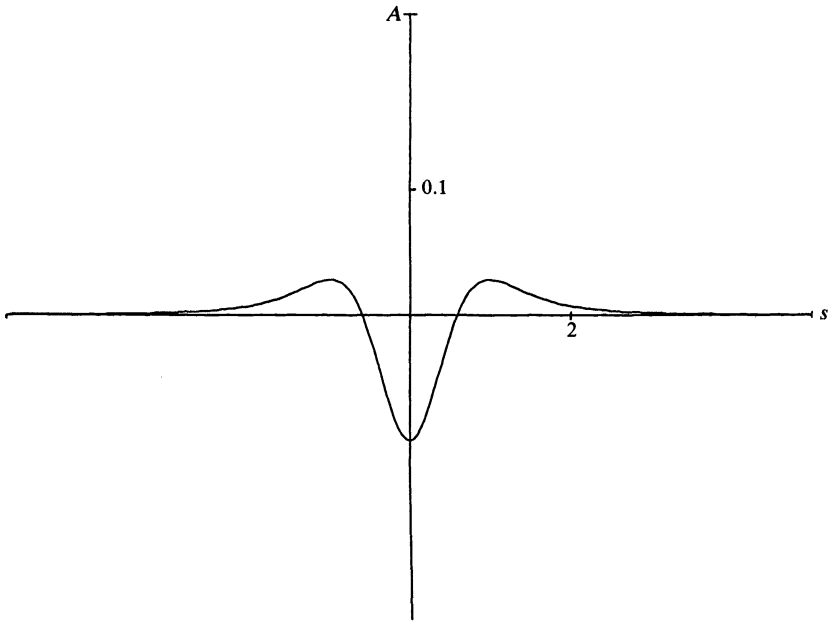


FIG. 5. $\sigma = 2$; $\beta = 1.25$; $L(-\infty) = 1$; $L(+\infty) = 1$. *Two values for* s_0: $s_0 \simeq +0.63$; $s_0 \simeq -0.63$.

For $\sigma = 2$ ($\alpha = 4$), Figs. 1-6 represent the function A for various values of β and $L(\pm\infty)$.

Acknowledgments. I am very grateful to Robert Magnus for showing me his work [9] prior to publication. From it I learned a great deal about the problem. I also thank David Chillingworth for his helpful remarks.

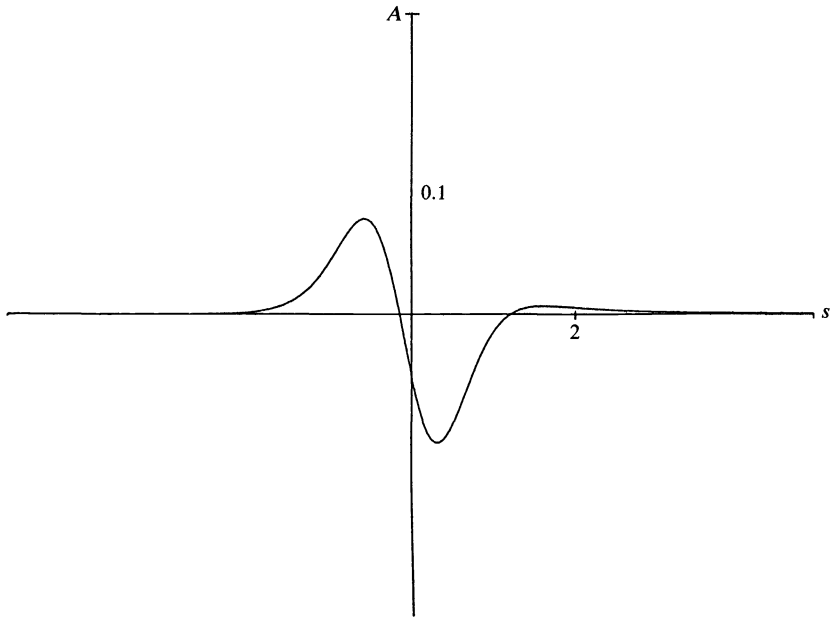


FIG. 6. $\sigma = 2$; $\beta = 1.25$; $L(-\infty) = 1$; $L(+\infty) = 0$. *Two values for* s_0: $s_0 \simeq +1.27$; $s_0 \simeq -0.13$.

REFERENCES

[1] R. CHIAPINELLI AND C. A. STUART, *Bifurcation when the linearisation has no eigenvalues*, J. Differential Equations, 30 (1978), pp. 269-307.

[2] S.-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, New York, 1982.

[3] S.-N. CHOW, J. K. HALE, AND J. MALLET-PARRET, *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37 (1980), pp. 351-373.

[4] W. A. COPPEL, *Stability and Asymptotic Behaviour of Differential Equations*, Heath, Boston, 1965.

[5] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321-340.

[6] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, Berlin, New York, 1983.

[7] T. KÜPPER AND D. REIMER, *Necessary and sufficient conditions for bifurcation from the continuous spectrum*, Nonlinear Anal. Theory Methods Appl., 3 (1979), pp. 555-561.

[8] R. J. MAGNUS, *On the asymptotic properties of solutions to a differential equation in a case of bifurcation without eigenvalues*, Proc. Roy. Soc. Edinburgh Sect. A, 104 (1986), pp. 137-160.

[9] ———, *On perturbations of a translationally invariant differential equation*, preprint.

[10] C. A. STUART, *Bifurcation pour des problèmes de Dirichlet et de Neumann sans valeurs propres*, C.R. Acad. Sci. Paris Sér. I. Math., 288A (1979), pp. 761-764.

[11] ———, *Bifurcation for Neumann problems without eigenvalues*, J. Differential Equations, 36 (1980), pp. 391-407.

[12] ———, *Bifurcation for Dirichlet problems without eigenvalues*, Proc. London Math. Soc., 45 (1982), pp. 169-192.

[13] ———, *Bifurcation in* $L^p(\mathbb{R})$ *for a semilinear equation*, J. Differential Equations, 64 (1986), pp. 294-316.

[14] ———, *A global branch of solutions to a semilinear equation on an unbounded interval*, Proc. Roy. Soc. Edinburgh Sect. A, 101 (1985), pp. 273-282.

[15] ———, *Bifurcation from the continuous spectrum in* $L^p(\mathbb{R})$, in Bifurcation: Analysis, Algorithms, Applications, Birkhäuser, Basel, 1987, pp. 306-318.

[16] ———, *Bifurcation in* $L^p(\mathbb{R})$ *for a semilinear elliptic equation*, Proc. London Math. Soc., 57 (1988), pp. 511-541.

[17] J. F. TOLAND, *Global bifurcation for Neumann problems without eigenvalues*, J. Differential Equations, 44 (1982), pp. 82-110.

[18] ———, *Positive solutions of nonlinear elliptic equations—existence and non-existence of solutions with radial symmetry in* $L^p(\mathbb{R}^N)$, Trans. Amer. Math. Soc., 282 (1984), pp. 335-354.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1172-1185, September 1989

009

EQUIVALENCE OF DIFFERENTIAL OPERATORS*

NIKY KAMRAN† AND PETER J. OLVER‡

Abstract. Two versions of the equivalence problem—determining when two second-order differential operators on the line are the same under a change of variables—are solved completely using the Cartan method of equivalence.

Key words. differential operator, Cartan equivalence method, invariant, symmetry, Lie algebra

AMS(MOS) subject classifications. 47E05, 58A15, 34B25

1. Introduction. The basic equivalence problem to be treated here is to determine when two second-order differential operators on the real line can be transformed into each other by an appropriate change of variables. There are two different possible interpretations of the notion of equivalence, depending on whether we wish to preserve the differential expression corresponding to the operator or the Lie bracket between operators. In this paper we treat both versions of the equivalence problem for second-order operators on the line. The problems here are related to the more general equivalence problem for second-order ordinary differential equations [4], [5], [11], but are specialized by linearity. We employ the equivalence method of Cartan, which gives necessary and sufficient conditions for equivalence. Although the simplest of the possible equivalence problems arising in the study of differential operators, these problems provide a good illustration of the power and ease of use of Cartan's equivalence method, which offers a straightforward algorithm for solving these and other equivalence problems important in applications. Extensions to higher-order or higher-dimensional operators can be readily done using the methods of this paper, although the intervening calculations will, as a rule, become much more complicated. In the proof of the theorem, we assume that the reader has a basic familiarity with the Cartan equivalence method as explained, for instance, in [2], [3], [6], and [7], although the reader can certainly understand the final results without all the intervening machinery.

This paper originated in answer to a question raised by Levine [9], who asked when a differential operator can be expressed as a bilinear combination of first-order differential operators that generate a finite-dimensional Lie algebra. This problem has applications to scattering theory in molecular dynamics and quantum chemistry. Indeed, there are now a number of well-established methods for dealing with such operators, where the calculation of eigenvalues, spectra, and dynamics is considerably simplified. The companion paper [8] applies the results of this paper to solving Levine's problem completely.

2. Equivalence problems for differential operators. Consider a second-order differential operator

$$\mathscr{D} = f(x)D^2 + g(x)D + h(x), \tag{2.1}$$

* Received by the editors June 6, 1988; accepted for publication October 31, 1988.

† School of Mathematics, Institute for Advanced Study, Princeton, New Jersey 08540; Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. The work of this author was partly supported by grants from the National Science Foundation and the Natural Sciences and Engineering Research Council of Canada.

‡ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was partly supported by National Science Foundation grant DMS 86-02004.

where f, g, h, are analytic functions of the real variable $x \in \mathbb{R}$, and $D = d/dx$. If we apply $\mathscr{D}$ to a scalar-valued function $u(x)$, we obtain the expression

$$\mathscr{D}[u] = fu'' + gu' + hu. \tag{2.2}$$

In particular, we can look at the linear, homogeneous second-order ordinary differential equation $\mathscr{D}[u] = 0$, or the eigenvalue problem $\mathscr{D}[u] = \lambda u$, or the Schrödinger equation $u_t = i\mathscr{D}[u]$, in which $\mathscr{D}$ plays the role of the Hamiltonian.

We will be concerned with the problem of when two such differential operators can be mapped into each other by an appropriate change of coordinates. It turns out that two natural classes of transformations can be employed to change the differential operator. Clearly, as far as the independent and dependent variables are concerned, the appropriate pseudogroup consists of the fiber-preserving transformations that are linear in the fiber variable u:

$$\bar{x} = \varphi(x), \qquad \bar{u} = \psi(x)u. \tag{2.3}$$

The total derivative operators are related by the chain rule formula[1]

$$\bar{D} = \frac{1}{\varphi'(x)} D. \tag{2.4}$$

In the first of our two equivalence problems, we identify the two differential expressions $\bar{\mathscr{D}}[\bar{u}] = \mathscr{D}[u]$ (cf. (2.2)), where

$$\bar{\mathscr{D}} = \bar{f}(\bar{x})\bar{D}^2 + \bar{g}(\bar{x})\bar{D} + \bar{h}(\bar{x})$$

is another second-order differential operator. The explicit formulae for the new coefficient functions $\bar{f}, \bar{g}, \bar{h}$, in terms of the original coefficients f, g, h of $\mathscr{D}$, can be determined using the transformation rule

$$\bar{\mathscr{D}} = \mathscr{D} \cdot \frac{1}{\psi(x)}, \tag{2.5}$$

together with the chain rule (2.4). The first of our equivalence problems for differential operators then amounts to determining conditions on the two differential operators such that there exists a transformation (2.3) that maps one to the other according to (2.5).

The transformation rule (2.5) has the disadvantage of not preserving either the eigenvalue problem or the Schrödinger equation associated with the operator. For instance, $\mathscr{D}[u] = \lambda u$ does not imply $\bar{\mathscr{D}}[\bar{u}] = \lambda \bar{u}$, since we are missing a factor of $\psi(x)$. To rectify this situation, we need to premultiply by $\psi(x)$ and use the alternative transformation rule

$$\bar{\mathscr{D}} = \psi(x) \cdot \mathscr{D} \cdot \frac{1}{\psi(x)}. \tag{2.6}$$

This transformation rule leads to slightly different formulae expressing the new coefficients $\bar{f}, \bar{g}, \bar{h}$, in terms of f, g, h. The transformations (2.6) enjoy the additional property of preserving the standard commutator Lie bracket $[\mathscr{D}, \mathscr{E}] = \mathscr{D} \cdot \mathscr{E} - \mathscr{E} \cdot \mathscr{D}$ between differential operators. The second equivalence problem is to determine conditions on two differential operators such that there exists a transformation (2.3) mapping one to the other according to (2.6).

[1] For simplicity, we will explicitly denote the pull-back maps only in the statements of the theorems.

We will solve both equivalence problems in this paper. Incidentally, if we try to combine all the transformations (2.3), (2.5), (2.6) (for different functions $\psi(x)$), we end up with the trivial result that all second-order differential operators are equivalent under this largest pseudogroup.

To apply Cartan's algorithm to either equivalence problem, we need to recast the transformation rules (2.3), and (2.5) or (2.6), in the language of differential forms. The appropriate space to work in will be the second jet space J^2, which has coordinates x, u, p, q. Here p represents the derivative u', and q the derivative u''. The immediate goal is to construct an appropriate coframe, or pointwise basis for the cotangent space T^*J^2, that will encode the relevant transformation rules for our problem(s). The first remark is that as long as $u \neq 0$, the pseudogroup of transformations (2.3) is uniquely prescribed by imposing the 1-form equations

$$d\bar{x} = \alpha\, dx, \tag{2.7}$$

$$\frac{d\bar{u}}{\bar{u}} = \frac{du}{u} + \beta\, dx. \tag{2.8}$$

Here α and β are functions J^2, whose precise form does not need to be specified in advance. Indeed, the first equation implies that $\bar{x} = \varphi(x)$, with $\alpha = \varphi'$, while the second necessarily requires the linearity of the transformation in u, so that $\bar{u} = \psi(x)u$, with $\beta = \psi'/\psi$. Note that the restriction to $u \neq 0$, which means that we are restricting our attention to either the positive or negative real u-axis, is inessential as far as the differential operator itself is concerned. (Indeed, analytic continuation will extend our results across the apparent singular subspace $u = 0$.)

For the derivative variables p and q to transform correctly, we need to preserve the *contact ideal* on J^2, which is the differential ideal generated by the pair of 1-forms $du - p\, dx$, $dp - q\, dx$. In general, a diffeomorphism $\Phi\colon J^2 \to J^2$ determines a contact transformation if and only if

$$d\bar{u} - \bar{p}\, d\bar{x} = \lambda(du - p\, dx), \tag{2.9}$$

$$d\bar{p} - \bar{q}\, d\bar{x} = \mu(du - p\, dx) + \nu(dp - q\, dx), \tag{2.10}$$

where λ, μ, ν, are functions on J^2. Equations (2.7)–(2.9) by themselves already constitute part of an overdetermined equivalence problem on J^2. There is an algorithm, due to Cartan, to reduce this to an equivalence problem of standard form, but in our case, we can do this by inspection. It is easy to see that the 1-form $(du - p\, dx)/u$ is invariant, so the identification

$$\frac{d\bar{u} - \bar{p}\, d\bar{x}}{\bar{u}} = \frac{du - p\, dx}{u} \tag{2.11}$$

can replace both (2.8) and (2.9). The reader can check that the 1-form identities (2.7), (2.10), (2.11), are equivalent to requiring that the transformation on J^2 be the prolongation of a point transformation of the special form (2.3), with the derivative variables p, q, transforming correctly. Therefore, we take as the first three elements of our eventual coframe the 1-forms

$$\omega_1 = dx, \quad \omega_2 = \frac{du - p\, dx}{u}, \quad \omega_3 = dp - q\, dx, \tag{2.12}$$

with the transformation rules

$$\bar{\omega}_1 = A\omega_1, \quad \bar{\omega}_2 = \omega_2, \quad \bar{\omega}_3 = B\omega_2 + C\omega_3, \quad A, C \neq 0,$$

where A, B, C, are functions on J^2. This much of the coframe is the same for each equivalence problem. To complete the coframe, we need to supplement these 1-forms with an additional 1-form, which will encode the action of the transformation rule (2.5) or (2.6) on the differential operator itself.

In both cases, there is an obvious invariant function for the problem. For the equivalence problem (2.5), the invariant is the differential expression (2.2), i.e.,

$$I(x, u, p, q) = \mathscr{D}[u] = f(x)q + g(x)p + h(x)u. \tag{2.13}$$

For the second problem (2.6), the invariant is slightly more complicated:

$$I(x, u, p, q) = \frac{\mathscr{D}[u]}{u} = \frac{f(x)q + g(x)p}{u} + h(x), \tag{2.14}$$

since we need to take care of the extra factor of ψ. In either case, we have $I(x, u, p, q) = \bar{I}(\bar{x}, \bar{u}, \bar{p}, \bar{q})$ under the identification (2.5) or (2.6). We therefore take our final 1-form to be the differential $\omega_4 = dI$, so that for the equivalence problem (2.5) we have

$$\omega_4 = f\,dq + g\,dp + h\,du + \{f'q + g'p + h'u\}\,dx, \tag{2.15}$$

whereas for the alternative problem (2.6) we take

$$\omega_4 = \frac{f}{u}\,dq + \frac{g}{u}\,dp - \frac{fq + gp}{u^2}\,du + \left\{\frac{f'q + g'p}{u} + h'\right\} dx. \tag{2.16}$$

In both cases, the four 1-forms $\omega_1, \omega_2, \omega_3, \omega_4$, provide a coframe on the subset

$$\Omega^* = \{(x, u, p, q) \in J^2 \,|\, u \neq 0 \text{ and } f(x) \neq 0\}. \tag{2.17}$$

From now on, we restrict our attention to a connected component $\Omega \subset \Omega^*$ of the subset (2.17); note that on such a component, the signs of $f(x)$ and u are fixed. We require only that the last coframe elements agree up to contact, i.e.,

$$\bar{\omega}_4 = D\omega_2 + E\omega_3 + \omega_4.$$

We therefore define the structure group

$$G = \left\{ \begin{pmatrix} A & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & B & C & 0 \\ 0 & D & E & 1 \end{pmatrix} : A, B, C, D, E \in \mathbb{R}, A \cdot C \neq 0 \right\},$$

which happens to be the same for both equivalence problems, even though the two coframes are different.

As a consequence of these preliminary considerations, we have successfully encoded our equivalence problem in terms of a coframe, and have shown the following.

PROPOSITION 2.1. *Let $\mathscr{D}$ and $\bar{\mathscr{D}}$ be second-order differential operators. Let $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ and $\{\bar{\omega}_1, \bar{\omega}_2, \bar{\omega}_3, \bar{\omega}_4\}$ be the corresponding coframes, on open subsets Ω and $\bar{\Omega}$ of the second jet space, given by* (2.12) *and* (2.15) *or* (2.16), *the choice of ω_4 and $\bar{\omega}_4$ depending on the equivalence problem under consideration. The differential operators are equivalent under the pseudogroup* (2.3) *according to the respective transformation rule* (2.5) *or* (2.6) *if and only if there is a diffeomorphism $\Phi: \Omega \to \bar{\Omega}$ that satisfies*

$$\Phi^*(\bar{\omega}_i) = \sum_{j=1}^{4} g_{ij}\omega_j, \qquad i = 1, \cdots, 4,$$

where $g = (g_{ij})$ is a G-valued function on J^2, and Φ^* denotes the pull-back map on differential forms.

To apply Cartan's algorithm for this equivalence problem, we must "lift" the coframes to the space $J^2 \times G$. The *lifted coframe* takes the form

$$\theta_1 = A\omega_1, \quad \theta_2 = \omega_2, \quad \theta_3 = B\omega_2 + C\omega_3, \quad \theta_4 = D\omega_2 + E\omega_3 + \omega_4, \tag{2.18}$$

where the coefficients A, B, C, D, E are now interpreted as coordinates in the structure group G. We then have the standard reformulation of the equivalence condition of Proposition 2.1.

PROPOSITION 2.2. *Under the setup of Proposition* 2.2, *two differential operators* $\mathscr{D}$ *and* $\bar{\mathscr{D}}$ *are equivalent if and only if there is a diffeomorphism* $\Psi: \Omega \times G \to \bar{\Omega} \times G$ *that commutes with the natural left action of* G, *and maps the appropriate lifted coframe elements to each other*:

$$\Psi^*(\bar{\theta}_i) = \theta_i, \qquad i = 1, \cdots, 4. \tag{2.19}$$

3. Solution of the first equivalence problem. To keep our presentation as short as possible, we will assume that the reader has some familiarity with the mechanics of Cartan's equivalence method as discussed, for instance, in [2], [3], and [6]. We will solve both equivalence problems, beginning with the setup (2.5), corresponding to the coframe element (2.15). The solutions are fairly similar intrinsically, although the parametric formulae differ. We present this case in detail, and briefly indicate how the other problem goes in the following section.

We begin with the lifted coframe (2.18), based on the base coframe (2.12), (2.15). The basic tool in Cartan's method is the invariance of the exterior derivative operation under smooth maps, so we begin by computing the differentials $d\theta_i$. They are found to have the form

$$\begin{aligned} d\theta_1 &= \alpha \wedge \theta_1 + \sigma_1, \\ d\theta_2 &= \sigma_2, \\ d\theta_3 &= \beta \wedge \theta_2 + \gamma \wedge \theta_3 + \sigma_3, \\ d\theta_4 &= \delta \wedge \theta_2 + \varepsilon \wedge \theta_3 + \sigma_4. \end{aligned}$$

Here α, β, γ, δ, ε, form a basis for the right-invariant 1-forms on the Lie group G, and the torsion terms take the form

$$\sigma_i = \sum_{j<k} \tau_{ijk} \theta_j \wedge \theta_k, \qquad i = 1, \cdots, 4.$$

In the absorption part of Cartan's process, we are allowed to replace each 1-form α, β, γ, δ, ε by an expression of the form $\alpha + \Sigma z_j \theta_j$, etc., where the functions z_j are chosen so as to make as many of the torsion coefficients τ_{ijk} vanish as possible. In the present setup we can readily "absorb" all the torsion components except

$$\tau_{212} = -\frac{B + Cp}{ACu}, \quad \tau_{213} = \frac{1}{ACu}, \quad \tau_{314} = \frac{C}{Af}, \quad \tau_{414} = \frac{E}{Af}.$$

These components are invariants of the problem. Since they depend on the group parameters, the next step in the process is to normalize them to as simple a form as possible through a suitable choice of the group parameters. There are two possible normalizations for these torsion components, depending on $\kappa_1 = \text{sign}\,(f(x) \cdot u)$, leading to two different branches for the equivalence problem. (However, as we remarked above, as far as the differential operator itself is concerned, the division into two branches is not essential, since we can always change the sign of u by restricting our attention to a different connected component Ω of the domain (2.17). It is nevertheless

convenient to retain the sign through our analysis.) We normalize the torsion components to 0, κ_1, 1, 0, respectively, by setting

$$A=\frac{\sigma_1}{\sqrt{|fu|}},\quad B=-\sigma_2 p\sqrt{\left|\frac{f}{u}\right|},\quad C=\sigma_2\sqrt{\left|\frac{f}{u}\right|},\quad E=0. \tag{3.1}$$

Here $\sigma_1=\pm 1$ is an undetermined sign that must be left ambiguous (even with the specification of κ_1), and $\sigma_2=\sigma_1\cdot(\operatorname{sign} f)$. (See [7] for a detailed discussion of these types of signs.) The normalizations (3.1) have the effect of reducing the original Lie group G to a one-parameter subgroup, with D the only remaining undetermined parameter.

In the second loop through the equivalence procedure, we substitute (3.1) into the formulas for the lifted coframe (2.18) and recompute the differentials. The unabsorbable torsion component $\tau_{413}=D$ can then be normalized to zero by setting $D=0$. Note that we could have avoided adding in contact terms in our definition of θ_4, since ω_4 turns out to already be an invariant form.

By normalizing the torsion components, we have managed to eliminate all of the group parameters. This has had the effect of (a) reducing the structure group to the identity, and (b) reducing the lifted invariant coframe to an invariant coframe on the base space J^2, known as an $\{e\}$-*structure* or local parallelism. The explicit formula for the invariant coframe comes from (2.18), (3.1), and we have

$$\begin{aligned}
\theta_1&=\frac{\sigma_1\,dx}{\sqrt{|fu|}},\\
\theta_2&=\frac{du-p\,dx}{u},\\
\theta_3&=\sigma_2\sqrt{\left|\frac{f}{u}\right|}\left\{-\frac{p}{u}(du-p\,dx)+(dp-q\,dx)\right\},\\
\theta_4&=f\,dq+g\,dp+h\,du+(f'q+g'p+h'u)\,dx.
\end{aligned} \tag{3.2}$$

Indeed, as the reader can check, these 1-forms do satisfy the invariance conditions

$$\bar\theta_i=\theta_i,\qquad i=1,2,3,4,$$

under the pseudogroup of transformations (2.3), (2.5). Applying the exterior derivative to the invariant coframe elements, and re-expressing the resulting 2-forms in terms of the coframe, we find that the structure equations for our problem take the form

$$\begin{aligned}
d\theta_1&=\tfrac12\theta_1\wedge\theta_2,\\
d\theta_2&=\kappa_1\theta_1\wedge\theta_3,\\
d\theta_3&=-I\theta_1\wedge\theta_2+\kappa_1 J\theta_1\wedge\theta_3+\theta_1\wedge\theta_4+\tfrac12\theta_2\wedge\theta_3,\\
d\theta_4&=0,
\end{aligned} \tag{3.3}$$

where

$$I=fq+gp+hu,\qquad J=\sigma_2\sqrt{\left|\frac{u}{f}\right|}\left(\frac12 f'-\frac32 p\frac{f}{u}-g\right). \tag{3.4}$$

Because the coframe is invariant, the functions I and J are the fundamental invariants of the problem. Note that we have recovered our original invariant (2.13) as one of the torsion components in the structure equations (3.3).

The *covariant derivatives* $F_{,\theta_i}$ of a function F with respect to the coframe (3.2) are defined by expressing the differential of F in terms of the invariant coframe:

$$dF = F_{,\theta_1}\theta_1 + F_{,\theta_2}\theta_2 + F_{,\theta_3}\theta_3 + F_{,\theta_4}\theta_4. \tag{3.5}$$

Explicitly,

$$\begin{aligned} F_{,\theta_1} &= \sigma_1\sqrt{|fu|}\,\hat{D}_x F, \\ F_{,\theta_2} &= uF_u + pF_p - \frac{pg+hu}{f}F_q, \\ F_{,\theta_3} &= \sigma_2\sqrt{\left|\frac{u}{f}\right|}\left(F_p - \frac{g}{f}F_q\right), \\ F_{,\theta_4} &= \frac{1}{f}F_q. \end{aligned} \tag{3.6}$$

Here $\hat{D}_x$ denotes the differential operator

$$\hat{D}_x = \frac{\partial}{\partial x} + p\frac{\partial}{\partial u} + q\frac{\partial}{\partial p} + R\frac{\partial}{\partial q}, \tag{3.7}$$

where

$$R = -\frac{gq + hp + f'q + g'p + h'u}{f}. \tag{3.8}$$

Note that if we differentiate the invariant equation $I = \text{constant}$ (which is the same as the ordinary differential equation $\mathscr{D}[u] = \text{constant}$) with respect to x and solve for the third-order derivative $r = u'''$, we recover (3.8). In this sense, $\hat{D}_x$ can be identified with the total derivative operator on J^2.

The covariant derivatives of any of the fundamental invariants (3.4), called the *derived invariants*, are also clearly invariants. Since the differentials of I and J are of the form

$$dI = \theta_4, \qquad dJ = \kappa_1 K\theta_1 + \tfrac{1}{2}J\theta_2 - \tfrac{3}{2}\kappa_1\theta_3,$$

the only independent derived invariant is

$$\begin{aligned} K &= \kappa_1 J_{,\theta_1} = \kappa_1\sigma_1\sqrt{|fu|}\,\hat{D}_x J \\ &= -\frac{3}{2}fq + \frac{3}{4}f\frac{p^2}{u} - \frac{1}{2}p(f'+g) + u\frac{2ff'' - f'^2 + 2f'g - 4fg'}{4f}. \end{aligned} \tag{3.9}$$

(Note that K does not have an ambiguous sign.) We can continue differentiating to deduce higher-order derived invariants; for example,

$$dK = L\theta_1 + (K - \tfrac{3}{2}I)\theta_2 - J\theta_3 - \tfrac{3}{2}\theta_4,$$

so we have one second-order derived invariant

$$L = K_{,\theta_1} = \kappa_1 J_{,\theta_1,\theta_1} = \sigma_1\sqrt{|fu|}\,\hat{D}_x K,$$

which we avoid writing out explicitly.

Given an $\{e\}$-structure as above, we define its *rank* to be the number of functionally independent invariants (including all possible derived invariants). The *order* of the $\{e\}$-structure is the highest-order derived invariant required to complete the independent set of invariants. According to the standard Jacobian criterion for functional

independence, the particular $\{e\}$-structure given by the coframe (3.3) will have rank 4 and order 2, provided $dI \wedge dJ \wedge dK \wedge dL \neq 0$, whereby I, J, K, L are a complete set of functionally independent invariants. Exceptional cases with lower rank (and lower order) can occur if this wedge product vanishes.

To investigate the structure of the invariants in more detail, we proceed as follows. Note first that since $f \cdot u \neq 0$, the invariants I and J are always functionally independent. We can eliminate p and q from the original equations (3.4):

$$p = \frac{f' - 2g}{3f} u - \frac{2}{3} \kappa_1 \sigma_2 J \sqrt{\left| \frac{u}{f} \right|},$$

$$q = -\frac{gp + hu - I}{f} = \frac{2g^2 - f'g - 3fh}{3f^2} u - \frac{2}{3} \kappa_1 \sigma_1 J \frac{g|u|^{1/2}}{|f|^{3/2}} + \frac{I}{f}.$$

Substituting these into (3.9), we find that we can write

$$K = a(x)u + \frac{1}{3} \kappa_1 J^2 - \frac{3}{2} I,$$

where

$$a = \frac{5gf' - 2f'^2 - 2g^2}{6f} + \frac{3}{2} h - g' + \frac{1}{2} f''. \tag{3.10}$$

If the function $a(x) \equiv 0$, then K is a function of I and J. An easy chain rule argument shows that in this case, besides I and J, there are no further independent derived invariants. Therefore, if $a \equiv 0$, the rank of the $\{e\}$-structure is 2, and the order is zero.

Otherwise, if a does not vanish identically, we can take

$$\tilde{K} = a(x)u$$

as a new independent invariant, and compute its derived invariant:

$$\tilde{L} = \tilde{K}_{,\theta_1} = \sigma_1 \sqrt{|fu|}\, \hat{D}_x \tilde{K} = \kappa_1 b(x) |\tilde{K}|^{3/2} - \frac{2}{3} J \tilde{K},$$

where

$$b(x) = \sigma_1 \frac{3a'f + af' - 2ag}{3\sqrt{|fa^3|}}. \tag{3.11}$$

Note that b is an invariant that depends only on x. If b is constant, then $I, J, \tilde{K}$ form a complete set of functionally independent invariants; the rank is 3 and the order 2. Otherwise, for b not constant, we have an $\{e\}$-structure of maximal rank, with $I, J, \tilde{K}, b$ comprising our four fundamental independent invariants. In this case, we complete the solution to the equivalence problem by computing one final derived invariant:

$$b_{,\theta_1} = \sigma_1 \sqrt{|fu|}\, b' = c(x) |\tilde{K}|^{1/2},$$

where

$$c(x) = \sigma_1 \sqrt{|f/a|}\, b'(x) \tag{3.12}$$

is also an invariant. In the case of an $\{e\}$-structure of maximal rank, the *determining function* F for our equivalence problem is prescribed by re-expressing c in terms of b, i.e., we write

$$c(x) = F[b(x)]. \tag{3.13}$$

Note that F may be a multiply-valued function.

Example 3.1. Let us consider the case of a simple operator

$$\mathscr{D} = D^2 + h(x) \tag{3.14}$$

of Sturm–Liouville type, i.e., $f = 1$, $g = 0$. Such operators play a key role in quantum mechanics, scattering theory, and the theory of the Korteweg–de Vries equation. Here

$$a(x) = \tfrac{3}{2}h(x),$$

so we have a structure of rank 2 if and only if $h \equiv 0$ and $\mathscr{D} = D^2$. As a result of our construction, we deduce that a second-order differential operator (2.1) is equivalent to the differential operator D^2 if and only if

$$3ff'' - 2f'^2 + 5f'g - 6fg' - 2g^2 + 9fh = 0. \tag{3.15}$$

Continuing, if $h \neq 0$, then

$$b = \pm\sqrt{\frac{3}{2}}\,\frac{h'}{|h|^{3/2}}.$$

Therefore b is constant if and only if $h(x) = (cx + d)^{-2}$, i.e., we have either a translate of the radial Laplace operator

$$\mathscr{D} = D^2 + \frac{k}{x^2}, \tag{3.16}$$

when $b = -\sqrt{6/k} \neq 0$, or $\mathscr{D} = D^2 + k$, when $b = 0$. Note that $k \neq 0$ can be scaled to 1.

Finally, in the case when b is not constant, then

$$c = \frac{hh'' - \frac{3}{2}h^2}{h^3}.$$

The determining function F will be found by writing

$$\frac{h''}{h^2} = \tilde{F}\left(\frac{h'}{h^{3/2}}\right), \tag{3.17}$$

from which

$$F(t) = \tilde{F}(t) - \tfrac{3}{2}t^2.$$

We note that the ordinary differential equation (3.17) can be solved explicitly by quadratures, owing to the presence of an obvious two-parameter symmetry group of translations in t and scalings; see [10].

In essence, the collection of all the invariants and their derived invariants will completely solve our equivalence problem, providing explicit necessary and sufficient conditions for two differential operators to be equivalent under a transformation (2.5). The following theorem is a consequence of general results on the equivalence of $\{e\}$-structures [6], [7].

THEOREM 3.2. *Let $\mathscr{D}$ and $\bar{\mathscr{D}}$ be real-analytic differential operators. Define the function $a(x)$ by* (3.10). *If $a \not\equiv 0$, then define the functions $b(x)$, $c(x)$ by* (3.11), (3.12). *Then $\mathscr{D}$ and $\bar{\mathscr{D}}$ are equivalent under a change of variables* (2.3), (2.5) *if and only if*:

(i) *$a \equiv \bar{a} \equiv 0$, in which case both $\mathscr{D}$ and $\bar{\mathscr{D}}$ are equivalent to the operator D^2; or*

(ii) *Both a and $\bar{a}$ do not vanish identically and $b = \bar{b}$ are constant, in which case both $\mathscr{D}$ and $\bar{\mathscr{D}}$ are equivalent to either the radial Laplace operator* (3.16) *or the operator $D^2 + 1$; or*

(iii) *Both a and $\bar{a}$ do not vanish identically, both b and $\bar{b}$ are not constant, the determining functions prescribed by* (3.13) *are identical, $F = \bar{F}$, and the equation $b(x) = \bar{b}(\bar{x})$ has a real solution branch.* (*For complex equivalence, the last statement is unnecessary.*)

The change of variables required to map one operator into the other is implicitly given as the solution to the equations

$$b(x)=\bar{b}(\bar{x}), \qquad \bar{a}(\bar{x})\bar{u}=a(x)u, \tag{3.18}$$

restricted so that the signs $\kappa_1=\operatorname{sign}(f\cdot u)$ and $\bar{\kappa}_1=\operatorname{sign}(\bar{f}\cdot\bar{u})$ agree: $\kappa_1=\bar{\kappa}_1$.

In fact, the connection with the operator (3.14) can be used to complete the solution to the equivalence problem.

THEOREM 3.3. *If $\mathscr{D}$ is a second-order differential operator, then there is a transformation* (2.3), (2.5) *taking $\mathscr{D}$ into the operator $\bar{D}^2+\frac{2}{3}a(\bar{x})$, where the potential $a(\bar{x})$ is given by the (relative) invariant* (3.10) *when $\bar{x}=\varphi(x)$. Moreover, two differential operators are equivalent if and only if their corresponding potentials differ by the rescaling and translation group $\bar{a}(\bar{x})=\lambda^2 a(\lambda\bar{x}+\delta)$.*

In other words, the equivalence class of a differential operator under (2.6) is completely determined by its potential; moreover, two potentials are equivalent if and only if they are rescaled translates of each other.

4. Solution of the second equivalence problem. In this case, we begin as before with the lifted coframe (2.18), now based on the base coframe (2.12), (2.16). In the first loop through the equivalence procedure, we are left with the unabsorbable torsion components

$$\tau_{212}=-\frac{B+Cp}{ACu}, \quad \tau_{213}=\frac{1}{ACu}, \quad \tau_{314}=\frac{C}{Afu}, \quad \tau_{414}=\frac{E}{Afu}.$$

Again, there are two branches, depending on $\kappa_1=\operatorname{sign} f$. Here the sign restriction is more essential than in the previous equivalence problem, since we cannot change the sign of f by a transformation of type (2.6). We normalize the torsion components to 0, κ_1, 1, 0, respectively, by setting

$$A=\frac{\sigma_1}{u\sqrt{|f|}}, \quad B=-\sigma_2 p\sqrt{|f|}, \quad C=\sigma_2\sqrt{|f|}, \quad E=0,$$

where σ_1 is an ambiguous sign, and $\sigma_2=\sigma_1\kappa_1$. In the second loop through the equivalence procedure, the unabsorbable torsion components $\tau_{413}=-\tau_{312}=D$ can both be normalized to zero by setting $D=0$. The final invariant coframe is now given by

$$\begin{aligned}
\theta_1&=\frac{\sigma_1\,dx}{u\sqrt{|f|}},\\
\theta_2&=\frac{du-p\,dx}{u},\\
\theta_3&=\sigma_2\sqrt{|f|}\left\{(dp-q\,dx)-\frac{p}{u}(du-p\,dx)\right\},\\
\theta_4&=\frac{f}{u}\,dq+\frac{g}{u}\,dp+\frac{fq+gp}{u^2}\,du+\left\{\frac{f'q+g'p}{u}+h'\right\}dx.
\end{aligned} \tag{4.1}$$

The structure equations take a slightly different form:

$$\begin{aligned}
d\theta_1&=0,\\
d\theta_2&=\kappa_1\theta_1\wedge\theta_3,\\
d\theta_3&=-2J\theta_1\wedge\theta_3+\theta_1\wedge\theta_4,\\
d\theta_4&=0,
\end{aligned} \tag{4.2}$$

where

$$J=\frac{\sigma_1}{4\sqrt{|f|}}\left(2g-f'+\frac{4pf}{u}\right) \tag{4.3}$$

is a fundamental invariant of the problem. Interestingly, the original invariant I given in (2.14) does *not* appear among the structure functions of the adapted coframe. Indeed, it is easy to see that it cannot appear even among the derived invariants of the structure functions, since only the derivative h' appears in the coframe (4.1), so it would be impossible to recover the function h, which appears in the expression for I, by differentiation. Thus, the invariant coframe (4.1) must be supplemented by the additional invariant I to effect the correct solution to the problem. Although we have come up with a nonstandard equivalence problem, Cartan himself was already aware of such possibilities. Indeed, in his original treatment of the equivalence method, he allows for the incorporation of additional function invariants into an equivalence problem, and, as he says, "Rien n'est changé à la solution..." [1, p. 725]. Here, we have one invariant provided by the structure functions, and one additional invariant, both of whose derived invariants must be taken into account when discussing the solution to the problem.

The *covariant derivatives* of a function F with respect to the coframe (4.1) are

$$\begin{aligned}
F_{,\theta_1}&=\sigma_1\sqrt{|f|}\,\hat{D}_x F,\\
F_{,\theta_2}&=uF_u+pF_p-\frac{fq+pg(1-u)}{f}F_q,\\
F_{,\theta_3}&=\frac{\sigma_2 u}{\sqrt{|f|}}\left(F_p-\frac{g}{f}F_q\right),\\
F_{,\theta_4}&=\frac{u}{f}F_q.
\end{aligned} \tag{4.4}$$

Here

$$\hat{D}_x=\frac{\partial}{\partial x}+p\frac{\partial}{\partial u}+q\frac{\partial}{\partial p}+R\frac{\partial}{\partial q}$$

is similar to the total derivative operator (3.7), but

$$R=\frac{fqp+gp^2}{u}-(gq+f'q+g'p+h'u) \tag{4.5}$$

is different. As in § 3, if we differentiate the equation $I=\text{constant}$ with respect to x and solve for the third order derivative $r=u'''$, then we recover the expression (4.5).

Since the differentials of I and J are of the form

$$dI=\theta_4, \qquad dJ=K\theta_1+\theta_3,$$

the only independent derived invariant is

$$\begin{aligned}
K=J_{,\theta_1}&=\sigma_1\sqrt{|f|}\,\hat{D}_x J\\
&=f\frac{qu-p^2}{u^2}+f'\frac{p}{2u}-\frac{2ff''-f'^2+2f'g-4fg'}{8f}.
\end{aligned} \tag{4.6}$$

Furthermore,

$$dK=L\theta_1-2J\theta_3+\theta_4,$$

so we have only one further second-order derived invariant:

$$L = K_{,\theta_1} = J_{,\theta_1,\theta_1} = \sigma_1 \sqrt{|f|}\, \hat{D}_x K.$$

Note that in the case of the transformation rule (2.6), there is always a one-parameter symmetry group of any differential operator, namely, the scaling $u \to \lambda u$. Since the invariants must respect this symmetry, there can be at most three functionally independent invariants. Thus, the rank of this $\{e\}$-structure can be at most 3, and this will happen when $dI \wedge dJ \wedge dK \neq 0$.

To investigate the structure of the invariants in more detail, we proceed as before. We solve (2.14), (4.3), for p and q:

$$p = u\left(\frac{\sigma_2 J}{\sqrt{|f|}} + \frac{f' - 2g}{4f}\right),$$

$$q = \frac{uI - uh - pg}{f} = u\left(\frac{I}{f} - \sigma_1 \frac{gJ}{|f|^{3/2}} + \frac{2g^2 - f'g - 4fh}{4f^2}\right).$$

Thus

$$K = -a(x) + I - \kappa_1 J^2, \tag{4.7}$$

where

$$a = \frac{8gf' - 3f'^2 - 4g^2}{16f} + h - \frac{1}{2}g' + \frac{1}{4}f''. \tag{4.8}$$

The only degenerate case is when a is constant, so the rank is 2 and the order zero. Otherwise the rank is 3, and we can take $a(x)$ as a new invariant. The final derived invariant is

$$b = a_{,\theta_1} = \sigma_1 \sqrt{|f|}\, a'. \tag{4.9}$$

The *determining function* is found by re-expressing b in terms of a:

$$b(x) = F[a(x)]. \tag{4.10}$$

Note that since b has the ambiguous sign σ_1, the determining function F is only prescribed up to an ambiguous $\pm$ sign. (Indeed, the orientation reversing change of variables $x \to -x$, $u \to u$, will change the sign of f.) One solution to this annoying complication is to replace the invariant b by its square $b^2 = \kappa_1 f a'^2$.

THEOREM 4.1. *Let $\mathscr{D}$ and $\bar{\mathscr{D}}$ be real-analytic differential operators. Define the functions $a(x)$, $b(x)$, by* (4.8), (4.9). *Then $\mathscr{D}$ and $\bar{\mathscr{D}}$ are equivalent under a change of variables* (2.3), (2.6) *if and only if the signs $\kappa_1 = \operatorname{sign}(f) = \bar{\kappa}_1 = \operatorname{sign}(\bar{f})$ agree, and either*:

(i) *$a = \bar{a} = k$ are constant, in which case both $\mathscr{D}$ and $\bar{\mathscr{D}}$ are equivalent to the operator $D^2 + k$; or*

(ii) *Both a and $\bar{a}$ are not constant, the determining functions prescribed by* (4.10) *are identical, $F = \bar{F}$, and the equation $a(x) = \bar{a}(\bar{x})$ has a real solution branch.*

Example 4.2. Let us consider the case of the operator $\mathscr{D} = D^2 + h(x)$. In this case, $a(x) = h(x)$; hence we have a structure of rank 2 if and only if h is constant. A differential operator (2.1) is equivalent to the differential operator $D^2 + k$ via (2.6) if and only if $a = k$, i.e.,

$$4ff'' - 3f'^2 + 8f'g - 8fg' - 4g^2 + 16fh = 16kf. \tag{4.11}$$

Otherwise, since $b = h'$, the determining function will be found by writing

$$h' = F(h). \tag{4.12}$$

For a fixed determining function F, the general solution of (4.12) are just the translates of h, i.e., $\tilde{h}(x) = h(x-\delta)$. We conclude that two operators of the form (3.14) are equivalent under the transformation group (2.6) if and only if their potentials are translates of each other.

Conversely, given a determining function F, we can always construct a corresponding potential $h(x)$ by solving the elementary first-order ordinary differential equation (4.12). We thus recover the classical result that a general second-order differential operator can always be transformed into an operator of the form (3.14).

THEOREM 4.3. *If $\mathscr{D}$ is a second-order differential operator, then there is a transformation* (2.6) *taking $\mathscr{D}$ into the operator $\pm\bar{D}^2 + a(\bar{x})$, where the potential $a(\bar{x})$ is given by the invariant* (4.8) *when $\bar{x} = \varphi(x)$, and the sign in front of $\bar{D}^2$ is determined by the sign of f, the coefficient of D^2 in $\mathscr{D}$. Moreover, two differential operators are equivalent if and only if their signs are the same and the corresponding potentials differ by a translation*: $\bar{a}(x) = a(x+\delta)$.

In other words, outside singular points where $f(x) = 0$, the equivalence class of a differential operator under (2.6) is completely determined by its potential and the sign of its leading coefficient; moreover, two potentials are equivalent if and only if they are translates of each other.

5. Symmetries of differential operators. We will call a group of transformations of the form (2.3) a symmetry group of the differential operator $\mathscr{D}$ if the corresponding transformation (2.5) or (2.6) leaves the operator unchanged. (This is more restrictive than the concept of a symmetry group of a differential equation [10].) It is interesting to see what the corresponding infinitesimal symmetry criteria are.

PROPOSITION 5.1. *Given a vector field $\mathbf{v} = \xi(x)(\partial/\partial x) + \eta(x)u(\partial/\partial u)$ that generates a one-parameter group of transformations of the form* (2.3) *on $\mathbb{R}^2$, define a corresponding first-order differential operator $\mathscr{V} = \xi(x)D + \eta(x)$. The group generated by $\mathbf{v}$ is a symmetry group of the differential operator $\mathscr{D}$ of the type* (2.5) *or of the type* (2.6) *if and only if the operator equation*

$$[\mathscr{V}, \mathscr{D}] + \eta \cdot \mathscr{D} = 0 \tag{5.1}$$

or, respectively,

$$[\mathscr{V}, \mathscr{D}] = 0 \tag{5.2}$$

holds.

In either case, the proof is straightforward. In the second case, the scaling vector field with $\eta = 1$ always generates a symmetry group.

Cartan's method gives us a complete handle on the symmetry group of an $\{e\}$-structure. If the structure has rank r and the underlying space has dimension n, then the symmetry group forms a Lie group of dimension $n-r$. For the differential operator equivalence problems, then, $n = 4$, and so the symmetry group will have dimension $4-r$, where r is the number of functionally independent invariants. This leads to the following results.

THEOREM 5.2. *Let $\mathscr{D}$ be a real-analytic differential operator, and consider the symmetries of the type* (2.5). *Define the functions $a(x)$, $b(x)$, $c(x)$ as in* § 3. *Then*:

(i) *$\mathscr{D}$ admits a two-parameter symmetry group if and only if $a \equiv 0$.*

(ii) *$\mathscr{D}$ admits a one-parameter symmetry group if and only if a does not vanish identically, and b is constant.*

(iii) *If b is not constant, then $\mathscr{D}$ can admit only a discrete symmetry group.*

Thus a differential operator (2.1) is equivalent to the differential operator D^2 if and only if it admits a two-parameter group of symmetries which is also equivalent to

the condition (3.15). The two-parameter symmetry group for the operator D^2 is, of course, generated by translations $(x, u) \to (x+k, u)$, and the scaling transformations $(x, u) \to (\lambda x, \lambda^2 u)$. Similarly, we have the result that the differential operator is equivalent to the radial Laplace operator (3.16) or the operator D^2+1 if and only if it admits a one-parameter group of symmetries. For the radial Laplace operator (3.16), the symmetry group is the scaling group $(x, u) \to (\lambda x, \lambda^2 u)$; for D^2+1 the translation group remains. Finally, for any other differential operator, the symmetry group is at most a discrete subgroup.

THEOREM 5.3. *Let $\mathcal{D}$ be a real-analytic differential operator, and consider the symmetries of the type* (2.6). *Let $a(x)$ be the corresponding potential* (4.8). *Then $\mathcal{D}$ always admits the one-parameter scaling symmetry group $(x, u) \to (x, \lambda u)$. Moreover, $\mathcal{D}$ admits a two-parameter symmetry group if and only if a is constant; otherwise there is only the possibility of additional discrete symmetries.*

See Hsu and Kamran [4] for more detailed information on the use of Cartan's equivalence method for determining the possible symmetry groups of general second-order ordinary differential equations.

REFERENCES

[1] E. CARTAN, *Les sous-groupes des groupes continus de transformations*, in Oeuvres complètes, Part II, Vol. 2, Gauthiers-Villars, Paris, 1953, pp. 719-856.

[2] ———, *Les problèmes d'équivalence*, in Oeuvres complètes, Part II, Vol. 2, Gauthiers-Villars, Paris, 1953, pp. 1311-1334.

[3] R. B. GARDNER, *Differential geometric methods interfacing control theory*, in Differential Geometric Control Theory, R. W. Brockett, R. Millman, and H. Sussman, eds., Birkhäuser, Boston, 1983, pp. 117-180.

[4] L. HSU AND N. KAMRAN, *Classification of second-order ordinary differential equations admitting Lie groups of fiber-preserving symmetries*, Proc. London Math. Soc., to appear.

[5] N. KAMRAN, K. G. LAMB, AND W. F. SHADWICK, *The local equivalence problem for $d^2y/dx^2 = F(x, y, dx/dy)$ and the Painlevé transcendants*, J. Differential Geometry, 22 (1985), pp. 139-150.

[6] N. KAMRAN, *Contributions to the study of the equivalence problem of Elie Cartan and its applications to partial and ordinary differential equations*, Mém. Cl. Sci. Acad. Roy. Belgique, to appear.

[7] N. KAMRAN AND P. J. OLVER, *Equivalence problems for first order Lagrangians on the line*, J. Differential Equations, to appear.

[8] ———, *Lie algebras of differential operators and Lie algebraic potentials*, J. Math. Anal. Appl., to appear.

[9] R. D. LEVINE, *Lie algebraic approach to molecular structure and dynamics*, in Mathematical Frontiers in Computational Chemical Physics, D. G. Truhlar, ed., IMA Vol. Math. Appl., 15, Springer-Verlag, Berlin, New York, 1988.

[10] P. J. OLVER, *Applications of Lie Groups to Differential Equations*, Graduate Texts in Math. 107, Springer-Verlag, Berlin, New York, 1986.

[11] M. A. TRESSE, *Détermination des invariants ponctuels de l'équation différentielle ordinaire du second ordre $y'' = \omega(x, y, y')$*, S. Hirzel, Leipzig, 1896.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1186–1207, September 1989

010

UNCONSTRAINED VARIATIONAL PRINCIPLES FOR EIGENVALUES OF REAL SYMMETRIC MATRICES*

GILES AUCHMUTY†

Abstract. Certain real-valued functions, whose critical points and critical values are related to the eigenvalues and eigenvectors of a real symmetric matrix, are described and analyzed. These functions, in general, are smooth and bounded below. Variational principles for finding various specific eigenvalues and eigenvectors of the matrix A are described. These problems have a Morse theory. They may be written as the difference of two convex functions, so there are also natural dual problems that include the classical constrained variational principles for eigenvalues and eigenvectors.

Key words. eigenvalues, real symmetric matrices, unconstrained optimization, variational principles

AMS(MOS) subject classifications. primary 49G05; secondary 15A18

1. Introduction. This paper describes and analyzes some unconstrained variational principles for the eigenvalues and eigenvectors of a real symmetric matrix. Extensions to weighted eigenproblems and to finding singular values and singular vectors of general real matrices are given.

For almost a century now, variational methods for finding or estimating eigenvalues of symmetric matrices have been synonymous with variants of Rayleigh's principle. They are constrained optimization problems and there is a large literature on such principles and their theory. For recent summaries see Chatelin [4] or Parlett [5].

Here we describe some families of unconstrained variational principles for finding various eigenvalues and eigenvectors of a real symmetric matrix A. As described in § 2, we consider smooth functions E that are the sum of a function of $\|x\|^2$ and a function of $\langle Ax, x\rangle$, with $\|x\|$ being the Euclidean norm. Then the nonzero critical points of E arise at eigenvectors of A of specific norm. The norm of the critical point, and the critical value of E, are functions of the corresponding eigenvalue. The Hessian of E at a critical point has the same eigenvectors as A, so we may compute its (Morse) index and describe the topological type of the critical point.

Much of this paper is devoted to analyzing some specific functions E that are bounded below on $\mathbb{R}^n$ and are minimized, or have critical points, at specific eigenvectors of A. Thus in § 3, a variational principle for the largest eigenvalue of a positive definite symmetric matrix is described and analyzed. It is used to obtain upper and lower bounds on this eigenvalue, and a deflation theory for finding the second, third, and subsequent eigenvalues is developed.

In § 4, we analyze certain families of variational principles including some that are applicable to weighted eigenvalue problems. These results are applied in § 5 to principles for finding the smallest or largest eigenvalues, or the eigenvalues of a symmetric matrix A closest to a preassigned number μ.

Some different classes of functions are described in § 6. They have the property that their only nonzero critical points arise at eigenvectors of A corresponding to positive eigenvalues of A. This property provides some results applicable in stability theory, and also enables us to give variational criteria for determining whether A has any eigenvalues in a given interval and for estimating these eigenvalues.

* Received by the editors February 3, 1988; accepted for publication (in revised form) November 29, 1988.

† Department of Mathematics, University of Houston, Houston, Texas 77004. This research was partially supported by National Science Foundation grant DMS-8701886 and by the Air Force Office of Scientific Research.

These principles are extended in § 7 to finding singular values and singular vectors of general real matrices, while § 8 describes some invariance and symmetry properties of these functionals.

All of these functions either have a specific, finite number of critical points, or else the set of critical points is infinite but has a finite number of bounded components. The functions analyzed can all be written as the difference of two convex functions. Hence we can use nonconvex duality theory as in [1] or [7] to describe various dual problems. Some of these are described in § 9; they may be either constrained or unconstrained variational principles.

Some of the analysis described here may be extended to provide variational principles for eigenvalues of compact linear operators on a Hilbert space or for self-adjoint linear elliptic eigenproblems as in [2] and [3]. Much of this analysis, however, is specific to finite-dimensional problems and has been developed with a view towards the numerical computation of eigenvalues.

2. General functionals. Throughout this paper. A will be a real symmetric $n \times n$ matrix with real eigenvalues $\lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_2 \leqq \lambda_1$. We shall use only real arithmetic and the Euclidean norm and inner products

$$\|x\| = \|x\|_2 = \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2}, \qquad \langle x, y \rangle = \sum_{j=1}^{n} x_j y_j$$

for vectors x, y in $\mathbb{R}^n$.

When B is an $m \times n$ real matrix, we shall write B^T for its transpose and $B = (b_{ij})$, with b_{ij} being the i, jth component. $\ker B = \{x \in \mathbb{R}^n : Bx = 0\}$ and I being the $n \times n$ identity matrix. When x, y are in $\mathbb{R}^n$, then $x \otimes y = (x_i y_j)$ is a rank 1, $n \times n$ matrix and the vector x is said to be normalized if $\|x\| = 1$. Any other terms from linear algebra that are not defined here should be taken as in Strang [6].

Our interest is in describing and analyzing certain smooth functions, defined on all of $\mathbb{R}^n$, whose extrema, or critical points, are related to the spectral properties of A. Most of the functions can be written in terms of $\|x\|$, $\|Ax\|$, and $\langle Ax, x \rangle$. In particular, many of the functions will have the form $E : \mathbb{R}^n \to \mathbb{R}$ defined by

$$E(x) = \Phi(\tfrac{1}{2}\|x\|^2) + \Psi(\tfrac{1}{2}\langle Ax, x \rangle) \tag{2.1}$$

where

(A1) $\Phi : [0, \infty) \to \mathbb{R}$ is continuous and twice continuously differentiable on $(0, \infty)$ with $\Phi'(s) \neq 0$ for $s \neq 0$;

(A2) $\Psi : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable with $\Psi(0) = 0$.

In this section we prove some general results about these functions and their critical points.

We write

$$\nabla E(x) = \left(\frac{\partial E}{\partial x_1}(x), \frac{\partial E}{\partial x_2}(x), \cdots, \frac{\partial E}{\partial x_n}(x) \right)^T,$$

$$D^2 E(x) = \left(\frac{\partial^2 E}{\partial x_i \, \partial x_j}(x) \right)$$

for the gradient and the Hessian, respectively, of E at a point x. For scalar functions, differentiation is denoted by a prime.

A point $\tilde{x}$ in $\mathbb{R}^n$ is said to be a critical point of E if either

(2.2) (i) E is differentiable at $\tilde{x}$ and $\nabla E(\tilde{x}) = 0$, or

(ii) E is not differentiable at $\tilde{x}$.

A critical value of E is the value of $E(\tilde{x})$ when $\tilde{x}$ is a critical point. E is minimized at $\tilde{x}$ provided

$$E(\tilde{x}) = \inf_{x \in \mathbb{R}^n} E(x). \tag{2.3}$$

A critical point $\tilde{x}$ of E is said to be nondegenerate if E is twice continuously differentiable at $\tilde{x}$ and $D^2E(\tilde{x})$ is a nonsingular matrix. When $\tilde{x}$ is a nondegenerate critical point, then its Morse index $i(\tilde{x})$ is the number of negative eigenvalues of $D^2E(\tilde{x})$. The behavior of E near a nondegenerate critical point may be classified by its Morse index. In particular, $\tilde{x}$ is a local minimum (maximum) if its Morse index is zero (or n); it is a saddle point if $1 \leqq i(\tilde{x}) \leqq n-1$.

The significant fact about the function (2.1) is that its nonzero critical points are certain specific eigenvectors of A. Moreover, the eigenvectors of the Hessian of E at critical points are precisely the eigenvectors of A. This may be summarized as follows.

THEOREM 1. *Let A be a real symmetric matrix with E defined by* (2.1). *Assume* (A1) *and* (A2) *hold; then E is twice continuously differentiable on $\mathbb{R}^n - \{0\}$ with*

$$\nabla E(x) = \Phi'(\tfrac{1}{2}\|x\|^2)x + \Psi'(\tfrac{1}{2}\langle Ax, x\rangle)Ax, \tag{2.4}$$

$$\begin{aligned} D^2E(x) = {} & \Phi'(\tfrac{1}{2}\|x\|^2)I + \Phi''(\tfrac{1}{2}\|x\|^2)x \otimes x \\ & + \Psi'(\tfrac{1}{2}\langle Ax, x\rangle)A + \Psi''(\tfrac{1}{2}\langle Ax, x\rangle)Ax \otimes Ax. \end{aligned} \tag{2.5}$$

When $\tilde{x}$ is a nonzero critical point of E, then

(i) *$\tilde{x}$ is an eigenvector of A corresponding to the eigenvalue*

$$\tilde{\lambda} = -\frac{\Phi'(\frac{1}{2}\|\tilde{x}\|^2)}{\Psi'(\frac{1}{2}\langle A\tilde{x}, \tilde{x}\rangle)}; \tag{2.6}$$

(ii) $\|\tilde{x}\| = \mu$, *where μ is a solution of*

$$\Phi'\left(\frac{\mu^2}{2}\right) + \tilde{\lambda}\Psi'\left(\tilde{\lambda}\frac{\mu^2}{2}\right) = 0; \tag{2.7}$$

(iii) *The eigenvectors of $D^2E(\tilde{x})$ are precisely the eigenvectors of A.*

Proof. From (A1)-(A2) we have that E is twice continuously differentiable, except perhaps at $x = 0$. When we use the chain rule, (2.4) and (2.5) follow.

When $\tilde{x}$ is a nonzero critical point of E, then (2.2) and (2.4) imply that

$$\Psi'(\tfrac{1}{2}\langle A\tilde{x}, \tilde{x}\rangle)A\tilde{x} = -\Phi'(\tfrac{1}{2}\|\tilde{x}\|^2)\tilde{x}. \tag{2.8}$$

If $\Psi'(\frac{1}{2}\langle A\tilde{x}, \tilde{x}\rangle) \neq 0$, then (2.6) holds. Suppose this does not hold; then $\Phi'(\frac{1}{2}\|\tilde{x}\|^2)\tilde{x} = 0$. From (A1), $\Phi'(s) \neq 0$ for $s \neq 0$; thus $\tilde{x} = 0$, which contradicts the assumption that $\tilde{x}$ is nonzero. Thus (i) holds.

Take the inner products of (2.8) with $\tilde{x}$. Then,

$$\Psi'\left(\frac{1}{2}\langle A\tilde{x}, \tilde{x}\rangle\right)\langle A\tilde{x}, \tilde{x}\rangle = -\Phi'\left(\frac{\mu^2}{2}\right)\mu^2$$

and we also have $\langle A\tilde{x}, \tilde{x}\rangle = \tilde{\lambda}\mu^2$, with $\tilde{\lambda}$ being the eigenvalue from (i). Hence (2.7) follows, as $\mu \neq 0$.

Let $\{e^{(j)}: 1 \leqq j \leqq n\}$ be an orthonormal set of eigenvectors of A with

$$Ae^{(j)} = \lambda_j e^{(j)}, \qquad 1 \leqq j \leqq n. \tag{2.9}$$

Since (i) and (ii) hold, then $\tilde{x} = \pm\mu e^{(k)}$ for some $\mu > 0$, $1 \leqq k \leqq n$. Then

$$D^2E(\tilde{x})e^{(j)} = [\Phi'(\tfrac{1}{2}\mu^2) + \lambda_j\Psi'(\tfrac{1}{2}\lambda_k\mu^2)]e^{(j)} + [\Phi''(\tfrac{1}{2}\mu^2) + \lambda_k^2\Psi''(\tfrac{1}{2}\lambda_k\mu^2)]\mu^2\delta_{jk}e^{(j)} \tag{2.10}$$

where δ_{jk} is the Kronecker delta.

When $j \neq k$, we have $D^2E(\tilde{x})e^{(j)} = \nu_j e^{(j)}$ with

$$\nu_j = \Phi'(\tfrac{1}{2}\mu^2) + \lambda_j\Psi'(\tfrac{1}{2}\lambda_k\mu^2).$$

When $j = k$, (2.10) implies that $D^2E(\tilde{x})e^{(k)} = \nu_k e^{(k)}$ with

$$\nu_k = \Phi'(\tfrac{1}{2}\mu^2) + \lambda_k\Psi'(\tfrac{1}{2}\lambda_k\mu^2) + [\Phi''(\tfrac{1}{2}\mu^2) + \lambda_k^2\Psi''(\tfrac{1}{2}\lambda_k\mu^2)]\mu^2.$$

Since $\{e^{(j)}: 1 \leqq j \leqq n\}$ is a basis of $\mathbb{R}^n$, we see that this describes all the eigenvectors of $D^2E(\tilde{x})$ and hence (iii) holds.

COROLLARY 1. *Take A, E as in Theorem* 1 *and suppose that $e^{(j)}$ is a normalized eigenvector of A corresponding to an eigenvalue λ_j. Then $\tilde{x} = \pm\mu e^{(j)}$ is a nonzero critical point of E if and only if $\mu \neq 0$ and*

$$\Phi'\left(\frac{1}{2}\mu^2\right) + \lambda_j\Psi'\left(\lambda_j\left(\frac{\mu^2}{2}\right)\right) = 0. \tag{2.11}$$

Proof. If $\tilde{x} = \pm\mu e^{(j)}$ is a nonzero critical point of E, then (2.8) implies (2.11). Conversely, if (2.11) holds for some $\mu \neq 0$ and some $1 \leqq j \leqq n$, then $\pm\mu e^{(j)}$ will be a solution of (2.8) and thus $\tilde{x} = \pm\mu e^{(j)}$ will be a nonzero critical point of E.

Usually the critical values of E depend only on the eigenvalues λ_j of A. We may regard (2.11) as an equation for μ_j^2 in terms of λ_j. Suppose it has only a finite number of solutions $\mu_k^2(\lambda_j)$, $1 \leqq k \leqq K(j)$. Then the corresponding values of $E(\tilde{x})$ are

$$\Phi(\tfrac{1}{2}\mu_k^2(\lambda_j)) + \Psi(\tfrac{1}{2}\lambda_j\mu_k^2(\lambda_j)), \qquad 1 \leqq k \leqq K(j)$$

and this may be regarded as a function of λ_j alone. In all the specific examples analyzed in this paper μ^2 is a single-valued function of λ and the critical values of E are relatively easy to compute.

When B is a general $m \times n$ matrix, consider the function $H: \mathbb{R}^n \to \mathbb{R}$ defined by

$$H(x) = \Phi(\tfrac{1}{2}\|x\|^2) + \theta(\tfrac{1}{2}\|Bx\|^2) \tag{2.12}$$

where both Φ and θ obey (A1). This function can be written in the form (2.1) with $\theta = \Psi$ and $A = B^TB$. The critical points of H are related to the singular vectors of B.

Let $\ker B = \{x \in \mathbb{R}^n: Bx = 0\}$; then the analogue of Theorem 1 is as follows.

THEOREM 2. *Suppose B is a real $m \times n$ matrix and H is defined by* (2.12) *with Φ and θ obeying* (A1). *Then H is twice continuously differentiable on $\mathbb{R}^n - \ker B$ with*

$$\nabla H(x) = \Phi'(\tfrac{1}{2}\|x\|^2)x + \theta'(\tfrac{1}{2}\|Bx\|^2)B^TBx,$$

$$D^2H(x) = \Phi'(\tfrac{1}{2}\|x\|^2)I + \Phi''(\tfrac{1}{2}\|x\|^2)x \otimes x + \theta'(\tfrac{1}{2}\|Bx\|^2)B^TB + \theta''(\tfrac{1}{2}\|Bx\|^2)B^TBx \otimes B^TBx.$$

When $\tilde{x}$ is a critical point of H in $\mathbb{R}^n - \ker B$, then

(i) *$\tilde{x}$ is a singular vector of B corresponding to the singular value*

$$\tilde{\nu} = \left[\frac{\Phi'(\tfrac{1}{2}\|\tilde{x}\|^2)}{\theta'(\tfrac{1}{2}\|B\tilde{x}\|^2)}\right]^{1/2}.$$

(ii) *$\|\tilde{x}\| = \mu$, where μ is a solution of*

$$\Phi'(\tfrac{1}{2}\mu^2) + \tilde{\nu}^2\theta'(\tfrac{1}{2}\tilde{\nu}^2\mu^2) = 0.$$

(iii) *The eigenvectors of $D^2H(\tilde{x})$ are the singular vectors of A.*

Proof. This follows from Theorem 1 as $\|Bx\|^2 = \langle B^TBx, x\rangle$ for all x in $\mathbb{R}^n$, and the (right) singular vectors of B are the eigenvectors of A.

In the following sections, we look at variational principles associated with functions of the form E or H. Essentially we shall be interested in describing functions E or H that have specific minima. Generally the minima of E (or H) arise only at specific eigenvectors (or singular vectors) of A (or B) and we choose the functions so as to find the smallest, the largest, or some other special eigenvalue of A. A function $f:\mathbb{R}^n \to \mathbb{R}$ is said to be coercive on $\mathbb{R}^n$ provided

$$\lim_{\|x\|\to\infty} \frac{f(x)}{\|x\|} = \infty.$$

3. Variational principles for the largest eigenvalues of positive definite matrices. Throughout this section, A will be a positive definite, symmetric $n \times n$ matrix with eigenvalues $0 < \lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_2 \leqq \lambda_1$. Here we shall describe some functions of the form (2.1) whose minima provide information on the largest eigenvalues and eigenvectors of A.

Consider the function $E_q:\mathbb{R}^n \to \mathbb{R}$ defined by

$$E_q(x) = \frac{1}{2}\|x\|^2 - \frac{1}{q}\langle Ax, x\rangle^{q/2} \tag{3.1}$$

where $1 \leqq q < 2$. This function has the form (2.1) with $\Phi(s) = s$, $\Psi(s) = q^{-1}(2s)^{q/2}$. Both Φ and Ψ obey (A1). When $q = 1$, this E_q is proportional to the function g in equation 8.3 of [1], where it was derived as a dual variational problem to Rayleigh's principle. The variational properties of E_q may be summarized as follows.

THEOREM 3. *Let A be a positive definite, symmetric matrix and E_q be defined by* (3.1) *with $1 \leqq q < 2$. Then*

(i) *E_q is coercive on $\mathbb{R}^n$ with*

$$\inf_{x\in\mathbb{R}^n} E_q(x) = \frac{-1}{2\gamma}\lambda_1^{\gamma} \tag{3.2}$$

where

$$\gamma = \frac{q}{(2-q)}. \tag{3.3}$$

This is attained at $\pm\lambda_1^{\gamma/2}e^{(1)}$, where $e^{(1)}$ is a normalized eigenvector corresponding to the eigenvalue λ_1.

(ii) *The nonzero critical points of E_q are at $\pm\lambda_k^{\gamma/2}e^{(k)}$, where $e^{(k)}$ is a normalized eigenvector of A corresponding to the eigenvalue λ_k.*

(iii) *The critical values of E_q are $\{-\lambda_k^{\gamma}/2\gamma;\ 1 \leqq k \leqq n\} \cup \{0\}$.*

(iv) *$\tilde{x} = \pm\lambda_k^{\gamma/2}e^{(k)}$ is a nondegenerate critical point of E_q if and only if λ_k is a simple eigenvalue of A. In this case, the Morse index of $\tilde{x}$ is $k-1$.*

Proof. For any x, we have $\langle Ax, x\rangle \leqq \lambda_1\|x\|^2$, so

$$E_q(x) \geqq \frac{1}{2}\|x\|^2 - \frac{1}{q}\lambda_1^{q/2}\|x\|^q.$$

Thus E_q is coercive when $1 \leqq q < 2$.

Since E_q is continuous and coercive on $\mathbb{R}^n$, it attains a finite infimum. E_q is also continuously differentiable on $\mathbb{R}^n - \{0\}$ with

$$\nabla E_q(x) = x - \langle Ax, x\rangle^{(q-2)/2}Ax. \tag{3.4}$$

Thus the nonzero critical points of E_q obey

$$Ax = \sigma(x)x \tag{3.5}$$

where $\sigma(x) = \langle Ax, x\rangle^{(2-q)/2}$. If $\tilde{x}$ is a nonzero critical point of E_q then it is an eigenvector of A corresponding to the eigenvalue $\tilde{\lambda} = \sigma(\tilde{x})$. Take the inner products of (3.5) with $\tilde{x}$; then

$$\langle A\tilde{x}, \tilde{x}\rangle = \sigma(\tilde{x})\|\tilde{x}\|^2,$$

or

$$\tilde{\lambda}^{\gamma} = \|\tilde{x}\|^2 \quad \text{where } \gamma = \frac{q}{2-q}. \tag{3.6}$$

Thus (ii) follows. Moreover,

$$E_q(\tilde{x}) = \left(\frac{1}{2} - \frac{1}{q}\right)\tilde{\lambda}^{\gamma} = -\frac{1}{2\gamma}\tilde{\lambda}^{\gamma} \tag{3.7}$$

on substituting back into (3.1). Thus (iii) holds and

$$\inf_{x\in\mathbb{R}^n} E_q(x) = \min_{1\leqq j\leqq n} -\frac{1}{2\gamma}\lambda_j^{\gamma} = -\frac{1}{2\gamma}\lambda_1^{\gamma}$$

as claimed in (i).

Differentiating (3.4) we have, when $x \neq 0$,

$$D^2E_q(x) = I - \langle Ax, x\rangle^{(q-2)/2}A + (2-q)\frac{Ax \otimes Ax}{\langle Ax, x\rangle^{2-(q/2)}}. \tag{3.8}$$

Let $\{e^{(1)}, \cdots, e^{(n)}\}$ be an orthonormal set of eigenvectors of A obeying (2.9) and suppose

$$\tilde{x} = \pm\lambda_k^{\gamma/2}e^{(k)}. \tag{3.9}$$

Then $D^2E_q(\tilde{x})e^{(j)} = (1-(\lambda_j/\lambda_k))e^{(j)} + (2-q)\delta_{jk}e^{(j)}$ as $\langle A\tilde{x}, \tilde{x}\rangle = \lambda_k^{2/(2-q)}$. Thus $D^2E_q(\tilde{x})$ is nonsingular if and only if λ_k is a simple eigenvalue of A. When this holds, we see that the number of negative eigenvalues of $D^2E_q(\tilde{x})$ is precisely $k-1$ when (3.9) holds as claimed in (iv).

This result has a number of interesting corollaries. The first is related to the number of critical points of E_q.

COROLLARY 1. *Let A, E_q be as in Theorem 3 and define*

$$C_j = \{x \in \mathbb{R}^n : \|x\|^2 = \lambda_j^{\gamma} \text{ and } Ax = \lambda_j x\} \tag{3.10}$$

for $1 \leqq j \leqq n$. Then the set of nonzero critical points of E_q is $\bigcup_{j=1}^n C_j$.

Moreover,

(a) *If λ_j is a simple eigenvalue of A, then C_j consists of exactly two points;*

(b) *If $\lambda_j = \lambda_{j+1} = \cdots = \lambda_{j+m_j}$ is an eigenvalue of multiplicity $m_j > 1$, then $C_j = C_{j+1} = \cdots = C_{j+m_j}$ is isomorphic to an $(m_j - 1)$-dimensional sphere and each point in this set is a degenerate critical point of E_q.*

Hence the set of all nonzero critical points of E_q has a finite number of bounded, connected components. When A has n distinct eigenvalues, it has exactly $2n$ elements.

Proof. (a) This description of the nonzero critical points of E_q follows from (ii) of Theorem 3. If λ_j is a simple eigenvalue of A, then

$$C_j = \{\pm\lambda_j^{\gamma/2}e^{(j)}\}$$

where $e^{(j)}$ is a normalized eigenvector of A corresponding to the eigenvalue λ_j.

(b) When λ_j is an eigenvalue of A of multiplicity $m_j \geqq 2$, then let $\{e^{(j)}, e^{(j+1)}, \cdots, e^{(j+m_j-1)}\}$ be an orthonormal basis of the eigenspace. Then

$$C_j = \left\{ x \in \mathbb{R}^n : x = \sum_{k=1}^{m_j} c_k e^{(j+k-1)} \text{ and } \sum_{k=1}^{m_j} c_k^2 = \lambda_j^\gamma \right\}.$$

This is diffeomorphic to a sphere of dimension $m_j - 1$, and from (iv) of Theorem 3, each of these critical points is degenerate.

The other claims now follow directly.

This function provides direct, lower bounds on λ_1. For any x, we have

$$\lambda_1^\gamma \geqq -2\gamma E_q(x). \tag{3.11}$$

We often would like to obtain upper bounds on λ_1. This can be done by using property (ii) of Theorem 3 and restricting the domain. Since E_q is coercive, we have that for any $R \geqq 0$

$$\alpha_R = \inf_{\|x\| \geqq R} E_q(x) \tag{3.12}$$

is well defined and finite, and this infimum is attained.

THEOREM 4. *Take A, E_q as in Theorem 3 and suppose that α_R is defined by* (3.12). *If $\alpha_R > -R^2/2\gamma$, then $\lambda_1 < R^{2/\gamma}$.*

Proof. When $R^{2/\gamma} \leqq \lambda_1$, then we have that $x = \pm \lambda_1^{\gamma/2} e^{(1)}$ lies in $A_R = \{x \in \mathbb{R}^n : \|x\| \geqq R\}$. Hence $\alpha_R = \alpha_0 = -\lambda_1^\gamma/2\gamma \leqq -R^2/2\gamma$.

When $\alpha_R > -(R^2/2\gamma)$, then $R > \lambda_1^{\gamma/2}$, as $\hat{x}$ is not in A_R. Thus $\lambda_1 < R^{2/\gamma}$ as required.

When $q = 1$, Theorem 3 reduces, essentially, to Theorem 8.1 of [1]. As q increases from 1 to 2, $\gamma(q) = q/(2-q)$ increases from 1 to infinity and the infimum in (3.2) increases to zero when $\lambda_1 \leqq 1$. For all values of λ_1, the expression $\lambda_1^\gamma/2\gamma$ is convex in γ and when $\lambda_1 > 1$, this goes to infinity as q approaches 2.

Part (iv) of Theorem 3 says that there is a good Morse theory for this function. When $q = 1$, this theory may be regarded as a dual theory to the Courant–Weyl minimax theory for eigenvalues and eigenvectors (see § 5 of [1]). We might ask whether there are modified versions of (3.1) that could provide variational characterizations of the second, third, or other eigenvalues of A.

Let $\{e^{(1)}, e^{(2)}, \cdots, e^{(n)}\}$ be an orthonormal set of eigenvectors of A so that (2.9) holds. Let $1 \leqq k \leqq n-1$ and define

$$V_k = \{x \in \mathbb{R}^n : \langle x, e^{(j)} \rangle = 0 \text{ for } 1 \leqq j \leqq k\}. \tag{3.13}$$

When we restrict E_q to V_k we have the following result.

THEOREM 5. *Let A be a positive definite, symmetric matrix, and let E_q and V_k be defined by* (3.1) *and* (3.13), *respectively. Then*

(i) *E_q is coercive on V_k with*

$$\inf_{x \in V_k} E_q(x) = -\frac{\lambda_{k+1}^\gamma}{2\gamma} \tag{3.14}$$

where γ is given by (3.3). *This is attained on $C_{k+1} \cap V_k$.*

(ii) *The nonzero critical points of E_q on V_k are $\lambda_j^{\gamma/2} \tilde{e}$, where $\tilde{e}$ is a normalized eigenvector of A corresponding to some eigenvalue λ_j of A with $j \geqq k+1$.*

(iii) *The critical values of E_q on V_k are $\{-\lambda_j^\gamma/2\gamma : k+1 \leqq j \leqq n\}$.*

Proof. Since E_q is coercive on $\mathbb{R}^n$, it is coercive on V_k. At a critical point of E_q on V_k, standard Lagrange multiplier theorems imply that

$$\nabla E_q(x) = \sum_{j=1}^{k} \mu_j e^{(j)} \tag{3.15}$$

for some real numbers $\mu_1, \cdots, \mu_k$. That is,

$$\tilde{x} - \langle A\tilde{x}, \tilde{x}\rangle^{(q-2)/2} A\tilde{x} = \sum_{j=1}^{k} \mu_j e^{(j)}.$$

Take inner products with $\tilde{x}$ and $e^{(l)}$ for $1 \leqq l \leqq k$. Since $\tilde{x}$ is in V_k, we find that

$$\|\tilde{x}\|^2 = \langle A\tilde{x}, \tilde{x}\rangle^{q/2},$$

$$\langle \tilde{x}, e^{(l)}\rangle - \lambda_l \langle \tilde{x}, e^{(l)}\rangle \|\tilde{x}\|^{-(2/\gamma)} = \mu_l \quad \text{for } 1 \leqq l \leqq k.$$

Thus $\mu_l = 0$ for $1 \leqq l \leqq k$ as $\langle \tilde{x}, e^{(l)}\rangle = 0$, and so $\tilde{x}$ is an eigenvector of A corresponding to an eigenvalue $\tilde{\lambda} = \langle A\tilde{x}, \tilde{x}\rangle^{(2-q)/2}$. The other results now follow just as in Theorem 3.

This variational principle of minimizing E_q on V_k for the $(k+1)$st eigenvalue λ_{k+1} is a constrained variational principle. It could be rewritten as an unconstrained variational principle by using the canonical projection Q_k of $\mathbb{R}^n$ onto V_k. Let $Q_k : \mathbb{R}^n \to V_k$ and $E_{qk} : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$Q_k x = x - \sum_{j=1}^{k} \langle x e^{(j)}\rangle, \tag{3.16}$$

$$E_{qk}(x) = \frac{1}{2}\|Q_k x\|^2 - \frac{1}{q}\langle AQ_k x, Q_k x\rangle^{q/2} \tag{3.17}$$

We see that

$$E_{qk}(x) = \begin{cases} E_q(x) & \text{for any } x \in V_k, \\ 0 & \text{for any } x \perp V_k, \end{cases}$$

and that

$$\inf_{x \in \mathbb{R}^n} E_{qk}(x) = -\frac{\lambda_{k+1}^{\gamma}}{2\gamma}. \tag{3.18}$$

This value is attained at $\hat{x} = \lambda_{k+1}^{\gamma/2}\tilde{e}$, where $\tilde{e}$ is a normalized eigenvector of A corresponding to the eigenvalue λ_{k+1} of A and lying in V_k. Moreover, if x is any point in $\mathbb{R}^n$ such that $Q_k x = \hat{x}$, we have $E_{qk}(x) = E_{qk}(\hat{x})$. So there is an affine subspace of minimizers of E_{qk} and E_{qk} is not coercive on $\mathbb{R}^n$.

4. Variational principles for other eigenvalues. A number of interesting, unconstrained variational principles for the eigenvalues and eigenvectors of a real symmetric matrix A can be described by the problem of extremizing the function $F_r : \mathbb{R}^n \to \mathbb{R}$ defined by

$$F_r(x) = \tfrac{1}{2}\langle r(A)x, x\rangle - \|x\|. \tag{4.1}$$

Here we require that

(A3) $r(A)$ be a rational function of A and a positive definite matrix.

The simplest examples are when $r(A)$ is linear, quadratic, or a polynomial in A. When $r(A)$ is linear in A, then F_r has the form (2.1). When $r(A) = c_1 A^2 + c_2 I$, then F_r has the form (2.12); both of these cases will be analyzed in § 5. The functional $r(A) = A^{-1}$, with A being positive definite, has been analyzed in § 8 of [1].

Henceforth we shall use the notation

$$E_j = \{x \in \mathbb{R}^n : \|x\| = 1 \text{ and } Ax = \lambda_j x\}, \qquad 1 \leqq j \leqq n.$$

Thus E_j is the set of all normalized eigenvectors of A corresponding to the eigenvalue λ_j. When λ_j is a simple eigenvalue, then $E_j = \{\pm e^{(j)}\}$ consists of exactly two points. Otherwise two or more of the sets E_j will be the same and each will be a bounded, closed, connected set. We shall also write

$$E_j^r = \{x \in \mathbb{R}^n : \|x\| = 1 \text{ and } r(A)x = r(\lambda_j)x\} \quad \text{for } 1 \leqq j \leqq n.$$

THEOREM 6. *Let A be a real symmetric matrix, and let F_r be defined by* (4.1) *with* (A3) *holding. Then*

(4.2) (i) *F_r is coercive on $\mathbb{R}^n$ and* $\inf_{x \in \mathbb{R}^n} F_r(x) = \min_{1 \leqq j \leqq n} -\dfrac{1}{2r(\lambda_j)}$.

(ii) *The nonzero critical points of F_r are $\tilde{x} = r(\lambda_k)^{-1} e^{(k)}$, where $e^{(k)}$ is in E_k^r and $1 \leqq k \leqq n$. F_r is minimized at the points $\hat{x} = r(\lambda_J)^{-1} e^{(J)}$, with $e^{(J)}$ in E_J^r and J being an integer that minimizes $r(\lambda_j)$, for $1 \leqq j \leqq n$.*

(iii) *The critical values of F_r are 0 and $-(2r(\lambda_k))^{-1}$ for $1 \leqq k \leqq n$.*

(iv) *$\tilde{x} = r(\lambda_k)^{-1} e^{(k)}$ is a nondegenerate critical point of F_r if and only if $r(\lambda_j) \neq r(\lambda_k)$ for all $j \neq k$, $1 \leqq j \leqq n$. The eigenvalues of $D^2 F_r(\tilde{x})$ are $\{r(\lambda_j) - r(\lambda_k) : 1 \leqq j \leqq n, j \neq k\} \cup \{r(\lambda_k)\}$.*

Proof. Since $r(A)$ is positive definite, there exists $c > 0$ such that

$$\langle r(A)x, x \rangle \geqq c\|x\|^2.$$

Hence $F_r(x) \geqq c(\|x\|^2 - \|x\|)/2$ for all x in $\mathbb{R}^n$.

Thus F_r is coercive on $\mathbb{R}^n$ and attains a finite infimum. F_r is continuously differentiable on $\mathbb{R}^n - \{0\}$ with

$$\nabla F_r(x) = r(A)x - \frac{x}{\|x\|}. \tag{4.3}$$

Thus the nonzero critical points $\tilde{x}$ of F_r are eigenvectors of $r(A)$ with $\|\tilde{x}\|$ being the inverse of the eigenvalues of $r(A)$. From the spectral mapping theorem we know that the eigenvalues of $r(A)$ are $\{r(\lambda_j) : 1 \leqq j \leqq n\}$. Thus the first part of (ii) holds.

Moreover, $\langle r(A)\tilde{x}, \tilde{x} \rangle = \|\tilde{x}\| = r(\lambda_j)^{-1}$, so $F_r(\tilde{x}) = -(1/2r(\lambda_j))$ for some j. Hence the rest of (ii), (iii), and (4.2) hold.

Differentiating (4.3) we find that, when $x \neq 0$,

$$\begin{aligned} D^2 F_r(x) &= r(A) - \frac{I}{\|x\|} + \frac{x \otimes x}{\|x\|^3} \\ &= r(A) - \|x\|^{-1} Q \end{aligned} \tag{4.4}$$

where $Q = I - P$ and P is the projection in the direction x;

$$Py = \frac{\langle x, y \rangle x}{\|x\|^2} \quad \text{for } y \text{ in } \mathbb{R}^n. \tag{4.5}$$

Thus when $\tilde{x} = r(\lambda_k)^{-1} e^{(k)}$ we have

$$D^2 F_r(\tilde{x}) = r(A) - r(\lambda_k)(I - P_k)$$

where $P_k y = \langle y, e^{(k)} \rangle e^{(k)}$ is the projection in the direction of this eigenvector. Thus

$$D^2 F_r(\tilde{x})\, e^{(j)} = \begin{cases} (r(\lambda_j) - r(\lambda_k)) e^{(j)} & \text{if } j \neq k, \\ r(\lambda_k) e^{(k)} & \text{if } j = k. \end{cases}$$

Hence (iv) of Theorem 6 holds, and for any particular $r(A)$ we can compute the Morse index of a nondegenerate critical point.

COROLLARY. *Suppose that A and F_r are as in Theorem 6. Then the set of all critical points of F_r has a finite number of bounded, closed components. It is finite if and only if $r(A)$ has n simple eigenvalues.*

Proof. This follows from (ii) of Theorem 6. We have that each E_j^r consists of exactly two points whenever $r(\lambda_j)$ is a simple eigenvalue of $r(A)$; otherwise it is an infinite set.

The results in this theorem could be generalized to the function where $\|x\|$ in (4.1) is replaced by $\|x\|^q/q$ with $1<q<2$, but this does not seem to provide any different information.

Sometimes we are interested in finding weighted, or generalized, eigenvalues and eigenvectors of A with respect to another positive definite symmetric matrix C. In this case we replace (4.1) with the function $J:\mathbb{R}^n\to\mathbb{R}$ defined by

$$J(x)=\tfrac{1}{2}\langle r(A)x, x\rangle-\langle Cx, x\rangle^{1/2}. \tag{4.6}$$

Now we assume that the weighted eigenproblem

$$Ax=\nu Cx \tag{4.7}$$

has real eigenvalues $\nu_n\leqq\nu_{n-1}\leqq\cdots\leqq\nu_1$, and we define the corresponding sets of normalized eigenvectors by

$$J_k=\{x\in\mathbb{R}^n\colon Ax=\nu_k Cx, \langle Cx, x\rangle=1\},$$

$$J_k^r=\{x\in\mathbb{R}^n\colon r(A)x=r(\nu_k)Cx, \langle Cx, x\rangle=1\}$$

where $1\leqq k\leqq n$.

The analogue of Theorem 6 is as follows.

THEOREM 7. *Let A, C be real symmetric matrices with C being positive definite and define J by* (4.6) *with* (A3) *holding. Then*

(4.8) (i) *J is coercive on $\mathbb{R}^n$ and* $\inf_{x\in\mathbb{R}^n} J(x)=\min_{1\leqq k\leqq n}-\dfrac{1}{2r(\nu_k)}$.

(ii) *The nonzero critical points of J are $\tilde{x}=r(\nu_k)^{-1}e^{(k)}$, where $e^{(k)}$ is in J_k^r and $1\leqq k\leqq n$. J is minimized at $\hat{x}=r(\nu_K)^{-1}e^{(K)}$, where K is the integer where $r(\nu_k)$ is minimized.*

(iii) *The critical values of J are 0 and $(-2r(\nu_k))^{-1}$ for $1\leqq k\leqq n$.*

(iv) *$\tilde{x}=r(\nu_k)^{-1}e^{(k)}$ is a nondegenerate critical point of J if and only if $r(\nu_j)\neq r(\nu_k)$ for all $j\neq k$, $1\leqq j\leqq n$. The eigenvalues of $D^2J(\tilde{x})$ with respect to C are $\{r(\nu_j)-r(\nu_k)\colon 1\leqq j\leqq n, j\neq k\}\cup\{r(\nu_k)\}$.*

Proof. This proof is similar to that of Theorem 6. Since C is bounded, we have $\langle Cx, x\rangle^{1/2}\leqq\|C\|^{1/2}\|x\|$, so

$$J(x)\geqq\frac{c}{2}\|x\|^2-\|C\|^{1/2}\|x\|$$

where c is as in Theorem 6. Hence J is coercive. Also, if $x\neq 0$,

$$\nabla J(x)=r(A)x-Cx/\sqrt{\langle Cx, x\rangle},$$

and now (i)-(iii) follow just as in Theorem 6. The Hessian at x is

$$D^2J(x)=r(A)-(C/\sqrt{\langle Cx, x\rangle})+\frac{Cx\otimes Cx}{\langle Cx, x\rangle^{3/2}}.$$

When $\tilde{x} = r(\nu_k)^{-1} e^{(k)}$ we have $\langle C\tilde{x}, \tilde{x}\rangle = \langle Ce^{(k)}, e^{(k)}\rangle / r(\nu_k)^2 = 1/r(\nu_k)^2$ so

$$D^2 J(\tilde{x}) e^{(j)} = r(A)e^{(j)} - r(\nu_k) Ce^{(j)} + r(\nu_k)^3 \langle C\tilde{x}, e^{(j)}\rangle C\tilde{x},$$

$$= \begin{cases} (r(\nu_j) - r(\nu_k))\, Ce^{(j)} & \text{if } j \neq k, \\ r(\nu_k) Ce^{(k)} & \text{if } j = k, \end{cases}$$

and thus (iv) holds.

In the remainder of this paper, we shall not pursue these generalized eigenvalue problems, but most of the results still hold mutatis mutandis. Essentially we replace $\|x\|$ by $\langle Cx, x\rangle^{1/2}$ and we obtain slightly more complicated expressions for derivatives and Hessians.

5. Examples of variational principles. By choosing $r(A)$ appropriately we can describe variational principles for certain different eigenvalues and eigenvectors of A. The first two examples have $r(A)$ linear or quadratic in A.

Example 1(a). Let $r(A) = A - \mu I$, where μ is chosen so that $\mu < \lambda_n$, with λ_n being the least eigenvalue of A. When A is positive definite we may take $\mu = 0$.

Define $F_1 : \mathbb{R}^n \to \mathbb{R}$ by

$$F_1(x) = \frac{1}{2}\langle Ax, x\rangle - \frac{\mu}{2}\|x\|^2 - \|x\|. \tag{5.1}$$

The following theorem shows that minimizing F_1 on $\mathbb{R}^n$ provides upper bounds on λ_n.

THEOREM 8. *Let A, μ, F_1 be as above. Then*

(i) $\inf_{x\in\mathbb{R}^n} F_1(x) = -1/2(\lambda_n - \mu)$ *and this is attained at* $(\lambda_n - \mu)^{-1} e^{(n)}$, *where* $e^{(n)}$ *is in* E_n.

(ii) *The nonzero critical points of* F_1 *are* $\{(\lambda_k - \mu)^{-1} e^{(k)} : e^{(k)} \in E_k\}$.

(iii) *The critical values of* F_1 *are* $\{-\frac{1}{2}(\lambda_k - \mu)^{-1} : 1 \leqq k \leqq n\} \cup \{0\}$.

(iv) $\tilde{x} = (\lambda_k - \mu)^{-1} e^{(k)}$ *is a nondegenerate critical point of* F_1 *if and only if* λ_k *is a simple eigenvalue of* A. *In this case the Morse index of* $\tilde{x}$ *is* $(n-k)$.

Proof. To obtain the proof just substitute $r(A) = A - \mu I$ and $r(\lambda_j) = \lambda_j - \mu$ in Theorem 6.

Example 1(b). We can also take $r(A) = \mu I - A$ if we choose μ so that $\mu > \lambda_1$, with λ_1 being the largest eigenvalue of A. Then define $F_2 : \mathbb{R}^n \to \mathbb{R}$ by

$$F_2(x) = \frac{\mu}{2}\|x\|^2 - \langle Ax, x\rangle - \|x\|. \tag{5.2}$$

In this case $\inf_{x\in\mathbb{R}^n} F_2(x) = -1/2(\mu - \lambda_1)$ and parts (ii)-(iv) of Theorem 8 hold similarly. This time the Morse index of $\tilde{x}$ is k.

Thus minimizing F_2 on $\mathbb{R}^n$ provides lower bounds on λ_1, and the critical points and critical values of F_1, F_2 provide information on all the eigenvalues and eigenvectors of A.

EXAMPLE 2. Let $r(A) = (A - \mu I)^2$, where μ is not an eigenvalue of A. Define $F(\mu, x)$ by

$$F(\mu, x) = \tfrac{1}{2}\|(A - \mu I)x\|^2 - \|x\|. \tag{5.3}$$

The following theorem shows that minimizing $F(\mu, \cdot)$ on $\mathbb{R}^n$ provides information on the eigenvalues of A close to μ.

THEOREM 9. *Let A be a real symmetric matrix with μ not an eigenvalue of A, and let $F(\mu, \cdot)$ be defined by* (5.3). *Then*

$$\text{(i)} \quad \inf_{x\in\mathbb{R}^n} F(\mu, x) = \min_{1\leqq j\leqq n} \frac{-1}{2(\lambda_j - \mu)^2}. \tag{5.4}$$

If J is an integer at which this minimum is attained, then $F(\mu, \cdot)$ is minimized at $(\lambda_J - \mu)^{-2} e^{(J)}$, with $e^{(J)}$ in $E_J^{(2)}$.

(ii) *The nonzero critical points of $F(\mu, \cdot)$ are $(\lambda_k - \mu)^{-2} e^{(k)}$ with $e^{(k)}$ in $E_k^{(2)}$.*

(iii) *The negative critical values of $F(\mu, \cdot)$ are $-\frac{1}{2}(\lambda_k - \mu)^2$ for $1 \leqq k \leqq n$.*

(iv) *$\tilde{x} = (\lambda_k - \mu)^{-2} e^{(k)}$ is a nondegenerate critical point of $F(\mu, \cdot)$ if and only if $(\lambda_k - \mu)^2 \neq (\lambda_j - \mu)^2$ for all $j \neq k$, $1 \leqq j \leqq n$.*

Here $E_k^{(2)} = \{x \in \mathbb{R}^n : \|x\| = 1$ and $(A - \mu I)^2 x = (\lambda_k - \mu)^2 x\}$.

Proof. To obtain the proof just substitute $(\lambda - \mu)^2$ for $r(\lambda)$ in Theorem 6.

We see that as μ varies, the minimum value of $F(\mu, \cdot)$ varies and provides information about all the eigenvalues of A. Let

$$\alpha(\mu) = \min_{1 \leqq j \leqq n} \frac{-1}{2(\lambda_j - \mu)^2} = \inf_{x \in \mathbb{R}^n} F(\mu, x). \tag{5.5}$$

We see that $\alpha(\mu) < 0$ for any real μ and it is singular at the eigenvalues of A. In fact, we have the following results.

LEMMA 5.1. *Let A be a real symmetric matrix with eigenvalues $\lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_2 \leqq \lambda_1$, and let $\alpha(\mu)$ be defined by (5.5). Then*

(i) *α is continuous and monotone decreasing on $(-\infty, \lambda_n)$.*

(ii) *α is continuous and monotone increasing on (λ_1, ∞).*

(5.6) (iii) *If $\alpha(\mu) \leqq -a$ with $a > 0$, then A has an eigenvalue in the interval $[\mu - d, \mu + d]$, where $d = (2a)^{-(1/2)}$.*

(iv) *A is positive definite if and only if $\alpha(0)$ is finite and $\alpha(\mu)$ is monotone decreasing on $(-\infty, 0)$.*

Proof. Parts (i), (ii), and (iv) follow directly from (5.5).

If $\alpha(\mu) \leqq -a$, then from (5.4) there is an integer J such that

$$(\lambda_J - \mu)^2 \leqq \frac{1}{2a}.$$

Thus (iii) follows.

Part (iii) of Lemma 5.1 enables us to find upper and lower bounds on the eigenvalues of A. Suppose we know that there is a unique eigenvalue λ_J of A in an interval (μ_1, μ_2). Then $\alpha(\mu)$ will be monotone decreasing on $[\lambda_J - d_1, \lambda_J)$ and monotone increasing on $(\lambda_J, \lambda_J + d_2]$, where $d_1 = \frac{1}{2}(\lambda_J - \lambda_P)$ and λ_P is the eigenvalue of A closest to λ_J but less than λ_J and $d_2 = \frac{1}{2}(\lambda_N - \lambda_J)$, where λ_N is the eigenvalue of A closest to λ_J but larger than λ_J.

Thus there are a number of simple, effective algorithms to improve these upper and lower bounds on λ_J using information about $\alpha(\mu)$.

Moreover, we can use this to find a maximization principle for λ_n and a minimization principle for λ_1.

Let $C(\mu) = \{x \in \mathbb{R}^n : F(\mu, x) < 0\}$ and let $\tilde{\lambda}_n$ be the second smallest eigenvalue of A, $\tilde{\lambda}_1$ being the second largest eigenvalue of A ($\tilde{\lambda}_n \neq \lambda_n$, $\tilde{\lambda}_1 \neq \lambda_1$). When $\lambda_1 = \lambda_n$ is the only eigenvalue of A we can take $\tilde{\lambda}_n = +\infty$, $\tilde{\lambda}_1 = -\infty$ in the following theorem.

THEOREM 10. *Let A be a real symmetric matrix, and let λ_1, $\tilde{\lambda}_1$, λ_n, $\tilde{\lambda}_n$, $C(\mu)$ be as above. If μ is in $(\lambda_n, \frac{1}{2}(\tilde{\lambda}_n + \lambda_n))$, then*

$$\lambda_n = \max_{x \in C(\mu)} [\mu - 1/\sqrt{2F(\mu, x)}], \tag{5.7}$$

while if μ is in $(\frac{1}{2}(\lambda_1 + \tilde{\lambda}_1, \lambda_1))$, then

$$\lambda_1 = \min_{x \in C(\mu)} [\mu + 1/\sqrt{2F(\mu, x)}]. \tag{5.8}$$

Proof. Suppose μ is in $(\lambda_n, \tilde{\lambda}_n)$; then from Theorem 9 we have

$$\inf_{x\in\mathbb{R}^n} F(\mu, x) = \frac{-1}{2(\lambda_n - \mu)^2} = F(\mu, \hat{x})$$

where $\hat{x} = (\lambda_n - \mu)^{-2} e^{(n)}$ and $e^{(n)}$ is in E_n. Thus for any x in $C(\mu)$ we have

$$\lambda_n \geqq \mu - 1/\sqrt{2F(\mu, x)}$$

and equality holds here when $x = \hat{x}$. Thus (5.7) holds.

The argument for (5.8) is similar.

Example 3. Consider the case where $r(A) = A^{-1}(A^2 + \mu^2 I)$ for some $\mu \geqq 0$. $r(A)$ will be positive definite provided A is. When A is not positive definite, choose $\gamma < \lambda_n$; then similar results will hold if we use $A - \gamma I$ in place of A. Define $F_3 : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}$ by

$$F_3(\mu, x) = \tfrac{1}{2}\langle (A + \mu^2 A^{-1})x, x\rangle - \|x\|. \tag{5.9}$$

Just as in the preceding example, minimizing $F_3(\mu, \cdot)$ on $\mathbb{R}^n$ provides information on the eigenvalues of A close to μ. In this case, however, the problem does not become unbounded below at the eigenvalues of A. The results may be summarized as follows.

THEOREM 11. *Let A be a symmetric positive definite matrix and F_3 be defined by* (5.9). *Then*

$$\text{(5.10)} \quad \text{(i)} \quad \inf_{x\in\mathbb{R}^n} F_3(\mu, x) = \min_{1\leqq j\leqq n} \frac{-\lambda_j}{2(\lambda_j^2 + \mu^2)}.$$

If J is an integer at which this minimum is attained, then $F_3(\mu, x)$ is minimized at $\lambda_J(\lambda_J^2 + \mu^2)^{-1} e^{(J)}$ with $e^{(J)}$ in $E_J^{(3)}$.

(ii) *The nonzero critical points of $f_3(\mu, \cdot)$ are $(\lambda_k/2)(\lambda_k^2 + \mu^2)^{-1} e^{(k)}$ with $e^{(k)}$ in $E_k^{(3)}$.*

(iii) *The critical values of $F_3(\mu, \cdot)$ are 0 and $-\lambda_k(\lambda_k^2 + \mu^2)^{-1}$ for $1 \leqq k \leqq n$.*

(iv) *$\tilde{x} = \lambda_k(\lambda_k^2 + \mu^2)^{-1} e^{(k)}$ is a nondegenerate critical point of $F_3(\mu, \cdot)$ if and only if λ_k is a simple eigenvalue of A and $\lambda_j \lambda_k \neq \mu^2$ for any $j \neq k$, $1 \leqq j \leqq n$.*

Here $E_k^{(3)} = \{x \in \mathbb{R}^n : \|x\| = 1$ and $Ax + \mu^2 A^{-1} x = (\lambda_k + \mu^2/\lambda_k)x\}$.

Proof. Use $r(\lambda) = \lambda/(\lambda^2 + \mu^2)$ in Theorem 6; then the result follows.

The function $r(\lambda) = \lambda/(\lambda^2 + \mu^2)$ is unimodal with $r(0) = r(\infty) = 0$ and is maximized at $\lambda = \mu$. Thus the minimizing J in (5.10) arises at an eigenvalue λ_J of A that is closest to μ in this special sense. By varying μ we can find all the distinct eigenvalues of A from information about the infima of $F_3(\mu, \cdot)$.

6. Variational principles for positive or negative eigenvalues. In the last three sections, the variational principles generally required some assumption involving positive definiteness of a matrix. Now consider the function $\mathcal{G}_p : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\mathcal{G}_p(x) = \frac{1}{p}\|x\|^p - \frac{1}{2}\langle Ax, x\rangle \tag{6.1}$$

with $2 < p < \infty$. This function has the form (2.1), with $\Psi(s) = -s$, $\Phi(s) = Cs^{p/2}$, and C a constant. For calculations we shall often take $p = 3$ or 4. When p increases to $+\infty$, the function $\mathcal{G}_p(x)$ converges pointwise to the function

$$\mathcal{G}_\infty(x) = \chi_1(x) - \tfrac{1}{2}\langle Ax, x\rangle \tag{6.2}$$

where

$$\chi_1(x) = \begin{cases} 0 & \text{if } \|x\| \leqq 1, \\ \infty & \text{otherwise.} \end{cases}$$

$\mathcal{G}_\infty$ arises in the convex analysis formulation of Rayleigh's principle (see § 8 of [1]).

The results on extremizing $\mathcal{G}_p$ may be summarized as follows.

THEOREM 12. *Suppose A is a real symmetric matrix and $\mathcal{G}_p$ is defined by* (6.1) *with $p > 2$. Then*

(6.3) (i) *$\mathcal{G}_p$ is coercive and $\inf_{x \in \mathbb{R}^n} \mathcal{G}_p(x) = (-1/2\nu)[\max(0, \lambda_1)]^\nu$, where $\nu = p/(p-2)$ and λ_1 is the largest eigenvalue of A.*

(ii) *The critical points of $\mathcal{G}_p$ occur at 0 and at $\tilde{x} = \lambda_j^{1/(p-2)} e^{(j)}$, where λ_j is a positive eigenvalue of A and $e^{(j)}$ is in E_j.*

(iii) *When $\lambda_1 > 0$, $\mathcal{G}_p$ is minimized at $\hat{x} = \lambda_1^{1/(p-2)} e^{(1)}$, where $e^{(1)} \in E_1$; if $\lambda_1 \leqq 0$ then $\mathcal{G}_p$ is minimized at $\hat{x} = 0$.*

(iv) *The critical values of $\mathcal{G}_p$ are $\{-\lambda_j^\nu/2\nu : \lambda_j > 0\} \cup \{0\}$.*

(v) *$\tilde{x} = \lambda_k^{1/(p-2)} e^{(k)}$ is a nondegenerate critical point of $\mathcal{G}_p$ if and only if λ_k is a simple eigenvalue of A. In this case the Morse index of $\tilde{x}$ is $(k-1)$. Zero is a nondegenerate critical point of $\mathcal{G}_p$ if and only if A is nonsingular.*

Proof. We have that $\mathcal{G}_p(x) \geqq \|x\|^p/p - \frac{1}{2}\|A\|\,\|x\|^2$ for all x in $\mathbb{R}^n$. Since $p > 2$, $\mathcal{G}_p$ is coercive and it attains its infimum on $\mathbb{R}^n$. Also

$$\nabla \mathcal{G}_p(x) = \|x\|^{p-2} x - Ax. \tag{6.4}$$

Thus if $\tilde{x}$ is a critical point of $\mathcal{G}_p$ then $\tilde{x} = 0$, or else $\tilde{x}$ is an eigenvector of A with eigenvalue $\tilde{\lambda} = \|\tilde{x}\|^{p-2} > 0$. That is, $\tilde{x} = 0$ or $\tilde{x} = \lambda_j^{1/(p-2)} e^{(j)}$ with $\lambda_j > 0$, $e^{(j)} \in E_j$. Thus (ii) holds and $\mathcal{G}_p(\tilde{x}) = 0$ or $-\lambda_j^\nu/2\nu$, where $\nu = p/(p-2)$. Thus (i), (iii), and (iv) follow.

Differentiating (6.4) we find that, when $x \neq 0$,

$$D^2 \mathcal{G}_p(x) = \|x\|^{p-2}[I + (p-2)P_x] - A \tag{6.5}$$

where $P_x y = \langle x, y \rangle x/\|x\|^2$ is the projection in the direction x. When $x = 0$, we have $D^2 \mathcal{G}_p(0) = -A$. When $\tilde{x} = \lambda_k^{1/(p-2)} e^{(k)}$ with $\lambda_k > 0$, we have

$$\begin{aligned} D^2 \mathcal{G}_p(\tilde{x}) e^{(j)} &= \lambda_k (e^{(j)} + (p-2)\delta_{jk} e^{(k)}) - A e^{(j)} \\ &= \begin{cases} (\lambda_k - \lambda_j) e^{(j)} & \text{if } j \neq k, \\ (p-2)\lambda_k e^{(k)} & \text{if } j = k. \end{cases} \end{aligned}$$

Thus (v) holds.

A symmetric matrix is said to be stable, or negative semidefinite, if all its eigenvalues are less than or equal to zero. A simple corollary of this theorem is the following.

COROLLARY 1. *A real symmetric matrix is stable if and only if $\inf_{x \in \mathbb{R}^n} \mathcal{G}_p(x) = 0$.*

Proof. The proof follows from (i) of Theorem 12.

To obtain the negative eigenvalues of A, we substitute $-A$ for A in (6.1). Alternatively, let

$$G_p(x) = \frac{1}{p}\|x\|^p + \frac{1}{2}\langle Ax, x \rangle \tag{6.6}$$

with $2 < p < \infty$, and consider the problem of extremizing G_p. The results may be summarized as follows.

COROLLARY 2. *Let A be a real symmetric matrix and G_p be defined by* (6.6) *with $p>2$. Then*

(i) *G_p is coercive and $\inf_{x\in\mathbb{R}^n} G_p(x)=(-1/2\nu)|\min(0,\lambda_n)|^\nu$ where $\nu=p/(p-2)$ and λ_n is the least eigenvalue of A.*

(ii) *The critical points of G_p occur at 0 and at $x=|\lambda_j|^{1/(p-2)}e^{(j)}$, where λ_j is a negative eigenvalue of A and $e^{(j)}\in E_j$.*

(iii) *The critical values of G_p are 0 and $-|\lambda_j|^\nu/2\nu$, where $\lambda_j<0$.*

(iv) *When $\lambda_n<0$, G_p is minimized at $|\lambda_n|^{1/(p-2)}e^{(n)}$, where $e^{(n)}\in E_n$. If $\lambda_n\geqq 0$, G_p is minimized at zero.*

(v) *Zero is a nondegenerate critical point of G_p if and only if A is nonsingular. When $\tilde{x}=|\lambda_k|^{1/(p-2)}e^{(k)}$ is a critical point of G_p, $\tilde{x}$ is nondegenerate if and only if λ_k is a simple eigenvalue of A. In this case the Morse index of $\tilde{x}$ is $n-k$.*

Proof. The proof is the same as the proof of Theorem 10.

COROLLARY 3. *Let A be a real symmetric matrix. Then A is positive semidefinite if and only if $G_p(x)\geqq 0$ for all x in $\mathbb{R}^n$.*

Proof. From Corollary 1 we see that $\lambda_n\geqq 0$ if and only if $\inf_{x\in\mathbb{R}^n} G_p(x)=0$; hence the result follows.

These variational principles may be modified to find the eigenvalues and eigenvectors of A lying in any preassigned interval. A functional whose critical values identify the eigenvalues of A in the interval (μ,∞) (or $(-\infty,\mu)$) is obtained by replacing A in (6.1) (or (6.6)) by $A-\mu I$.

To see if A has an eigenvalue in the interval (μ_1,μ_2) replace A in (6.1) by $(A-\mu_1 I)(\mu_2 I-A)$ and obtain

$$\mathcal{G}_p(\mu_1,\mu_2,x)=\frac{1}{p}\|x\|^p+\frac{1}{2}[\|Ax\|^2-(\mu_1+\mu_2)\langle Ax,x\rangle+\mu_1\mu_2\|x\|^2]. \tag{6.7}$$

Then we have the following corollary.

COROLLARY 4. *Let A be a real symmetric matrix, and let $\mathcal{G}_p(\mu_1,\mu_2,x)$ be defined by* (6.7). *Then A has an eigenvalue in the interval (μ_1,μ_2) if and only if*

$$\inf_{x\in\mathbb{R}^n}\mathcal{G}_p(\mu_1,\mu_2,x)=-\alpha(\mu_1,\mu_2)<0. \tag{6.8}$$

If $\alpha(\mu_1,\mu_2)>0$ then either λ_+ or λ_- is an eigenvalue of A, where

$$\lambda_\pm=\tfrac{1}{2}[(\mu_1+\mu_2)\pm d], \tag{6.9}$$

$$d^2=(\mu_2-\mu_1)^2-4(2\nu\alpha(\mu_1,\mu_2))^{1/\nu}. \tag{6.10}$$

If there is a $\tilde{c}=\mathcal{G}_p(\mu_1,\mu_2,\tilde{x})<0$ then A has an eigenvalue in the subinterval $[\mu_0-\delta,\mu_0+\delta]$, where $\mu_0=\frac{1}{2}(\mu_1+\mu_2)$ and

$$\delta^2=\tfrac{1}{4}(\mu_1-\mu_2)^2-(2\nu|\tilde{c}|)^{1/\nu}.$$

Proof. Equation (6.7) has the form (6.1) with $(A-\mu_1 I)(\mu_2 I-A)$ replacing A. Thus this infimum is negative if and only if $(A-\mu_1 I)(\mu_2 I-A)$ has a positive eigenvalue from (6.3). The eigenvalues of $(A-\mu_1 I)(\mu_2 I-A)$ are $(\lambda_j-\mu_1)(\mu_2-\lambda_j)$ for $1\leqq j\leqq n$, so (6.8) holds.

Let

$$\nu_1=\max_{1\leqq j\leqq n}(\lambda_j-\mu_1)(\mu_2-\lambda_j)>0$$

$$=(\lambda_J-\mu_1)(\mu_2-\lambda_J)\quad\text{for some } 1\leqq J\leqq n.$$

Then from (6.3), $(2\nu\alpha(\mu_1,\mu_2))^{1/\nu}=(\lambda_J-\mu_1)(\mu_2-\lambda_J)$. This is a quadratic equation for λ_J, whose solution is given by (6.9)-(6.10). More generally, if $\alpha(\mu_1,\mu_2)>|\tilde{c}|$, then we

have

$$(\lambda_J - \mu_1)(\mu_2 - \lambda_J) \geqq (2\nu|\tilde{c}|)^{1/\nu}$$

and thus we have that λ_J lies in $[\mu_0 - d, \mu_0 + d]$ as claimed.

In a similar manner, when μ is not an eigenvalue of A, we can find the eigenvalue closest to μ and greater than μ by substituting $(A-\mu I)^{-1}$ for A in (6.1).

Write

$$J_p(x) = \frac{1}{p}\|x\|^p - \frac{1}{2}\langle (A-\mu I)^{-1}x, x\rangle. \tag{6.11}$$

Then we have the following corollary.

COROLLARY 5. *Suppose A is real symmetric, μ is not an eigenvalue of A, and J_p is defined by* (6.11). *Then*

$$\inf_{x\in\mathbb{R}^n} J_p(x) = \begin{cases} \dfrac{-1}{2\nu}\left[\max\limits_{1\leqq j\leqq n}\left(\dfrac{1}{\lambda_j-\mu}\right)\right]^{\nu} & \text{if } \lambda_1 > \mu, \\ 0 \quad \text{if } \lambda_1 < \mu \end{cases} \tag{6.12}$$

with $\gamma = p/p-2$. When $\lambda_1 > \mu$, then this infimum is attained at $(\lambda_J - \mu)^{-1/(p-2)}e^{(J)}$, where λ_J is the eigenvalue of A closest to μ but larger than μ and $e^{(J)}$ is in E_J.

Proof. The eigenvalues of $(A-\mu I)^{-1}$ are $(\lambda_j - \mu)^{-1}$, and hence (6.12) follows from (6.3) and (6.11).

7. Variational principles for singular values. Suppose B is a general $m \times n$ real matrix. We can modify the preceding variational principles to describe functions whose critical points and critical values identify the singular values and singular vectors of B.

Throughout this section we assume B has singular values

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_p \geqq 0 \quad \text{with } p = \min(m, n). \tag{7.1}$$

Let $u^{(j)}$, $v^{(j)}$, respectively, be left and right normalized singular vectors of B corresponding to the singular value σ_j. Thus $\|u^{(j)}\| = \|v^{(j)}\| = 1$ and

$$Bv^{(j)} = \sigma_j u^{(j)}, \qquad B^T u^{(j)} = \sigma_j v^{(j)} \quad \text{for } 1 \leqq j \leqq p. \tag{7.2}$$

The functions to be extremized will all have the form (2.12). We shall look particularly at

$$H_1(x) = \tfrac{1}{2}\|Bx\|^2 - \|x\|, \tag{7.3}$$

$$H_2(x) = \tfrac{1}{2}\|x\|^2 - \|Bx\|, \tag{7.4}$$

$$H_p(x) = \frac{1}{p}\|x\|^p - \frac{1}{2}\|Bx\|^2 \quad \text{for } 2 < p < \infty. \tag{7.5}$$

Here H_1 generalizes $F(0, \cdot)$, where $F(\mu, \cdot)$ is defined by (5.3), H_2 has the form (3.1) with $q = 1$ and $A = B^TB$, and H_p is related to $\mathcal{G}_p$ defined by (6.1) with $A = B^TB$. Each of these functions provides information on the singular values and singular vectors of B.

First consider H_1. Its extremal properties may be summarized as follows.

THEOREM 13. *Let B be a real $m \times n$ matrix and H_1 be defined by* (7.3). *Then*

(i) *H_1 is coercive on $\mathbb{R}^n$ if and only if* rank $B = n$.

(ii) *If* rank $B = n$, *then* $\inf_{x\in\mathbb{R}^n} H_1(x) = -1/2\sigma_n^2$, (7.6)

(7.7) *while if* rank $B \leqq n-1$, *then* $\inf_{x\in\mathbb{R}^n} H_1(x) = -\infty$.

(iii) *H_1 is continuously differentiable on $\mathbb{R}^n - \{0\}$ and the nonzero critical points of H_1 are $\sigma_j^{-2} v^{(j)}$ with $v^{(j)}$ being a normalized right singular vector of B corresponding to σ_j.*

(iv) *The critical values of H_1 are $\{-1/2\sigma_j^2 : 1 \leqq j \leqq p\}$.*

(v) *$\tilde{x} = \sigma_k^{-1} v^{(k)}$ is a nondegenerate critical point of H_1 if and only if σ_k^2 is a simple, nonzero eigenvalue of B^TB. In this case the Morse index of $\tilde{x}$ is $n-k$.*

Proof. If H_1 is coercive on $\mathbb{R}^n$, then there exists $R > 0$ such that

$$\|x\| \geqq R \text{ implies } \|Bx\| \geqq 1.$$

Hence $\|Bx\| \geqq \|x\|/R$ and thus rank $B = n$.

Conversely if rank $B = n$, then there exists $c > 0$ such that

$$\|Bx\| \geqq c\|x\| \quad \text{for all } x \text{ in } \mathbb{R}^n,$$

and hence H_1 is coercive. Thus (i) holds.

Now if rank $B \leqq n-1$, then there is an $\tilde{x} \neq 0$ in $\mathbb{R}^n$ such that $B\tilde{x} = 0$. Then $H(t\tilde{x}) = -t\|\tilde{x}\|$ and, letting $t \to \infty$, we see that (7.7) holds.

Differentiating (7.3), we find that, for $x \neq 0$,

$$\nabla H_1(x) = B^TBx - \frac{x}{\|x\|}. \tag{7.8}$$

Hence if $\tilde{x}$ is a nonzero critical point of H_1 we have that $\tilde{x}$ is a right singular vector of B and $\|\tilde{x}\|^{-1} = \sigma_j^2$ gives the corresponding singular value. Thus (iii) holds and we have

$$\|B\tilde{x}\|^2 = \|\tilde{x}\| = \sigma_j^{-2}$$

on taking inner products with $\tilde{x}$. Hence $H_1(\tilde{x}) = -1/2\sigma_j^2$, so (iv) holds, and when rank $B = n$ then (7.6) holds.

Differentiating (7.8) we obtain

$$D^2H_1(x) = B^TB - \frac{1}{\|x\|}\left(I - \frac{x\otimes x}{\|x\|^2}\right),$$

and thus (v) holds just as in the proof of Theorem 6, with $r(\lambda_j)$ being replaced by the eigenvalues σ_j^2 of B^TB.

Theorem 13 indicates that H_1 will be most informative when rank $B = n$ and thus $m \geqq n$. The functionals for H_2 and H_p, however, are always bounded below. The results may be summarized as follows.

THEOREM 14. *Let B be a real $m \times n$ matrix, and let H_2 be defined by* (7.4). *Then*

(7.9) (i) *H_2 is coercive on $\mathbb{R}^n$ and $\inf_{x\in\mathbb{R}^n} H_2(x) = -\sigma_1^2/2$. This infimum is attained at $\sigma_1 v^{(1)}$, where $v^{(1)}$ is a normalized singular vector corresponding to the singular value σ_1.*

(ii) *H_2 is continuously differentiable on $\mathbb{R}^n - \ker B$ and its critical points are* ker B *together with $\{\sigma_j^2 v^{(j)} : v^{(j)}$ is a normalized right singular vector corresponding to σ_j, $1 \leqq j \leqq p\}$.*

(iii) *The critical values of H_2 are $\{-\sigma_j^2/2 : 1 \leqq j \leqq p\} \cup \{0\}$.*

(iv) *$\tilde{x} = \sigma_k v^{(k)}$, with $\sigma_k \neq 0$, is a nondegenerate critical point of H_2 if and only if σ_k^2 is a simple nonzero eigenvalue of B^TB. In this case its Morse index is $k-1$.*

Proof. We have $H_2(x) = \frac{1}{2}\|x\|^2 - (\langle B^TBx, x\rangle)^{1/2}$. Thus H_2 has the form (3.1) with $q = 1$ and $A = B^TB$. The eigenvalues λ_j of A are related to the singular values σ_j of B by

$$\lambda_j = \begin{cases} \sigma_j^2 & \text{for } 1 \leqq j \leqq p, \\ 0 & \text{if } j > p, \end{cases}$$

and the eigenvectors of A are the right singular vectors of B. The proof of this result now follows the pattern of the proof of Theorem 3 except that now A is only positive semidefinite.

THEOREM 15. *Let B be a real $m \times n$ matrix, and let H_p be defined by* (7.5) *with $p > 2$. Then*

(7.10) (i) *H_p is coercive on $\mathbb{R}^n$ and $\inf_{x \in \mathbb{R}^n} H_p(x) = -\sigma_1^{\eta p}/\eta p$, where $\eta = 2/(p-2)$. This is attained at $\sigma_1^{\eta} v^{(1)}$.*

(ii) *The critical points of H_p occur at 0 and at $\tilde{x} = \sigma_j^{\eta} v^{(j)}$, where σ_j is a positive singular value of B and $v^{(j)}$ is a corresponding normalized singular vector.*

(iii) *The critical values of H_p are $\{-\sigma_j^{p\eta}/p\eta : \sigma_j > 0\} \cup \{0\}$.*

(iv) *$\tilde{x} = \sigma_k^{\eta} v^{(k)}$ is a nondegenerate critical point of H_p if and only if σ_k^2 is a simple eigenvalue of $B^T B$. In this case, the Morse index of $\tilde{x}$ is $(k-1)$.*

Proof. This result is proven in the same manner as Theorem 12 with $B^T B$ in place of A and σ_j^2 in place of λ_j.

It is worth noting that minimizing H_2 or H_p for $p > 2$ provides estimates on σ_1, the largest singular value of B. Minimizing H_1 tells us whether rank B is n and, if that is so, it estimates σ_n. To find other singular values of B we could use functions such as (5.3), (5.9), or (6.7) with A replaced by $B^T B$ or $B^T B + I$, and then use the corresponding theorems to find the singular values closest to μ or in some interval. Similarly, results such as the corollary to Theorem 6 can be proven for each of these functions.

8. Invariance properties. A function $E : \mathbb{R}^n \to \mathbb{R}$ is said to be invariant under a linear transformation $L : \mathbb{R}^n \to \mathbb{R}^n$ provided

$$E(Lx) = E(x) \quad \text{for all } x \text{ in } \mathbb{R}^n. \tag{8.1}$$

If E is invariant under a nonsingular linear transformation L and $\tilde{x}$ is a critical point of E, then $L\tilde{x}$ will also be a critical point of E and it will have the same value as $E(\tilde{x})$ and the same Morse index if it is nondegenerate.

We observe that all the functions introduced so far are invariant under the linear transformation $L = -I$. Particular matrices A often have other symmetries. The following result is often very useful.

THEOREM 16. *Let A be a real symmetric $n \times n$ matrix and let U be a real orthogonal matrix that commutes with A. Then E_q in* § 3, *F_r in* § 4, *F, F_1, F_2, and F_3 in* § 5, *and the functions $\mathcal{G}_p$ and G_p in* § 6 *are invariant under U.*

Proof. All the norms have been 2-norms and U orthogonal implies that $\|Ux\| = \|x\|$. Also $\langle AUx, Ux\rangle = \langle U^{-1}AUx, x\rangle = \langle Ax, x\rangle$ as U commutes with A. Moreover, $\langle r(A)Ux, Ux\rangle = \langle r(A)x, x\rangle$ and $\|(A-\mu I)Ux\| = \|U(A-\mu I)x\| = \|(A-\mu I)x\|$ so all the functions listed above are invariant under U as claimed.

This theorem can be used to analyze symmetries in the sets of critical points of these functions.

9. Dual problems. All of the specific functions introduced in the preceding sections may be written as the difference of two convex functions. Thus we can use nonconvex duality theory, as developed in [1] or [7], to obtain various natural dual problems. As will be seen, many of these are well-known constrained variational problems.

Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is a function, then its conjugate convex (or polar) function $f^* : \mathbb{R}^n \to (-\infty, \infty]$ is defined by

$$f^*(y) = \sup_{x \in \mathbb{R}^n} [\langle x, y\rangle - f(x)].$$

To describe the dual variational principles we shall need the dual functions of f_1 and f_2, where

$$f_1(x) = \frac{1}{p}\|Ax\|^p, \tag{9.1}$$

$$f_2(x) = \frac{1}{p}\langle Ax, x\rangle^{p/2} \tag{9.2}$$

with $1 \leqq p < \infty$ and A real symmetric. A must be positive semidefinite in (9.2). To describe the polar functions of f_1, f_2 we shall need the indicator functional $\chi_K(x)$ of a closed, convex set K in $\mathbb{R}^n$ defined by

$$\chi_K(x) = \begin{cases} 0 & \text{if } x \in K, \\ \infty & \text{otherwise.} \end{cases} \tag{9.3}$$

LEMMA 9.1. *Suppose A is real symmetric and nonsingular and f_1 is defined by* (9.1). *Then*

(9.4) (i) *When* $1 < p < \infty$ $f_1^*(y) = \|A^{-1}y\|^q/q$ *with* $q = p/(p-1)$;

(9.5) (ii) *When* $p = 1$ $f_1^*(y) = \begin{cases} 0 & \text{if } \|A^{-1}y\| \leqq 1, \\ \infty & \text{otherwise.} \end{cases}$

Proof. (i) When $1 < p < \infty$ we have

$$f_1^*(y) = \sup_{x \in \mathbb{R}^n} \left[\langle x, y\rangle - \frac{1}{p}\|Ax\|^p \right].$$

The expression on the right-hand side is maximized when

$$y = \|Ax\|^{p-2}A^2x \tag{9.6}$$

as A is symmetric. Thus, $A^{-1}y = \|Ax\|^{p-2}Ax$ and $\|A^{-1}y\| = \|Ax\|^{p-1}$. Thus the solution $\hat{x}$ of (9.6) is given by

$$\hat{x} = \frac{A^{-2}y}{\|A^{-1}y\|^r} \quad \text{where } r = \frac{p-2}{p-1}.$$

If we substitute this in the expression for f_1^* we obtain (9.4).

(ii) When $p = 1$, extremality condition (9.6) becomes

$$y = \frac{A^2x}{\|Ax\|}.$$

Thus the maximizing $\hat{x}$ lies in the direction of $A^{-2}y$. Consider

$$\begin{aligned} \Phi(s) &= \langle sA^{-2}y, y\rangle - |s|\,\|A^{-1}y\| \\ &= s\|A^{-1}y\|^2 - |s|\,\|A^{-1}y\|. \end{aligned}$$

We see that $\Phi : \mathbb{R} \to \mathbb{R}$ obeys

$$\max_{-\infty < s < \infty} \Phi(s) = \begin{cases} 0 & \text{if } \|A^{-1}y\| \leqq 1, \\ \infty & \text{otherwise} \end{cases}$$

and thus (9.5) follows. □

It is worth noting that (9.4) implies that for A real symmetric and nonsingular,

$$\frac{1}{p}\|Ax\|^p + \frac{1}{q}\|A^{-1}y\|^q \geqq \langle x, y\rangle \tag{9.7}$$

for all x, y in $\mathbb{R}^n$ and where $1/p+1/q=1$. This is a form of Hölder's (or Young's) inequality. Also, (9.5) says that when $p=1$, $f_1^*(y)=\chi_{K_1}(y)$, where $K_1=\{y\in\mathbb{R}^n: \|A^{-1}y\|\leqq 1\}$.

LEMMA 9.2. *Let A be a positive semidefinite matrix, and let f_2 be defined by* (9.2). *Then*

$$\text{(9.8)}\quad \text{(i)}\quad \textit{When } 1<p<\infty \quad f_2^*(y)=\begin{cases} q^{-1}\langle A^+y, y\rangle^{q/2} & \textit{for } y\in R(A),\\ +\infty & \textit{otherwise};\end{cases}$$

$$\text{(9.9)}\quad \text{(ii)}\quad \textit{When } p=1 \quad f_2^*(y)=\chi_{K_2}(y).$$

Here $q=p/(p-1)$ is the conjugate index to p, A^+ is the Moore-Penrose inverse of A, $R(A)$ is the range of A and $K_2=\{y\in R(A): \langle A^+y, y\rangle\leqq 1\}$.

Proof. We have $f_2^*(y)=\sup_{x\in\mathbb{R}^n}[\langle x, y\rangle-\langle Ax, x\rangle^{p/2}/p]$.

Suppose (i) $1<p<\infty$; then the expression on the right is concave in x and will be maximized if and only if there is a solution $\hat{x}$ of

$$y=\langle Ax, x\rangle^{(p-2)/2}Ax. \tag{9.10}$$

When y is not in the range of A we have $f_2^*(y)=+\infty$. When y is in the range of A, and if $\tilde{x}$ is a solution of (9.10), we have that

$$\langle y, \tilde{x}\rangle=\langle A\tilde{x}, \tilde{x}\rangle^{p/2}$$

on taking inner products with $\tilde{x}$. Substituting, we have

$$f_2^*(y)=\frac{1}{q}\langle A\tilde{x}, \tilde{x}\rangle^{p/2} \quad \text{where}\ \frac{1}{p}+\frac{1}{q}=1. \tag{9.11}$$

Let $\{e^{(1)}, \cdots, e^{(m)}\}$ be an orthonormal set of eigenvectors of A that is a basis of $R(A)$ with $Ae^{(j)}=\lambda_j e^{(j)}$, $1\leqq j\leqq m$, $\lambda_j>0$. Assume $y=\sum_{j=1}^m d_j e^{(j)}$ and $\tilde{x}=\sum_{j=1}^m c_j e^{(j)}$. Then from (9.8), on taking inner products with $e^{(k)}$, we have

$$d_k=\left(\sum_{j=1}^m \lambda_j c_j^2\right)^{(p-2)/2}\lambda_k c_k=\alpha\lambda_k c_k$$

where $\alpha=(\sum_{j=1}^m \lambda_j c_j^2)^{(p-2)/2}>0$. Now from (9.9) we have

$$\sum_{k=1}^m c_k d_k=\frac{1}{\alpha}\sum_{k=1}^m \frac{d_k^2}{\lambda_k}=\alpha^{p/(p-2)} \quad \text{if } p\neq 2.$$

Thus

$$\alpha^{2(p-1)/(p-2)}=\sum_{k=1}^m \frac{d_k^2}{\lambda_k}$$

and so we have a unique positive solution for α if $p\neq 2$. When $p=2$, $\alpha=1$. Thus there is a unique $\tilde{x}$ in $R(A)$ that satisfies (9.10) and this is given by $\tilde{x}=\gamma A^+y$ for some constant γ, with A^+ being the Moore-Penrose inverse.

Now $\langle A^+y, y\rangle=\langle A\tilde{x}, \tilde{x}\rangle^{p-1}$ from (9.10), so (9.11) implies that for y in $R(A)$,

$$f_2^*(y)=\frac{1}{q}\langle A^+y, y\rangle^{p/(2(p-1))}$$

as required.

(ii) When $p=1$, the extremality condition (9.10) becomes

$$y=Ax/\sqrt{\langle Ax, x\rangle}.$$

When y is not in the range of A, $f_2^*(y) = +\infty$. When y is in the range of A, we can do another expansion in positive eigenvectors of A to show that $f_2^*(y) = \chi_{K_2}(y)$.

COROLLARY. *Let A be positive definite and let f_2 be defined by* (9.2). *Then*

(9.12) (i) *When* $p \neq 1$ $f_2^*(y) = \langle A^{-1}y, y\rangle^{q/2}/q$ with $q = p/(p-1)$;

(9.13) (ii) *When* $p = 1$ $f_2^*(y) = \chi_{K_2}(y)$ with $K_2 = \{y \in \mathbb{R}^n : \langle A^{-1}y, y\rangle \leqq 1\}$.

Proof. When A is nonsingular, A^+ becomes A^{-1}.

We are now in a position to describe the dual variational principles to the ones introduced in §§ 3–7. When $E : \mathbb{R}^n \to \mathbb{R}$ is a function of the form

$$E(x) = f_1(Lx) - f_2(x) \tag{9.14}$$

with f_1, f_2 being convex and lower semicontinuous on $\mathbb{R}^n$ and L an $n \times n$ matrix, then the dual problem is to minimize $\mathscr{E} : \mathbb{R}^n \to \bar{\mathbb{R}}$, where

$$\mathscr{E}(y) = f_2^*(L^T y) - f_1^*(y). \tag{9.15}$$

This is described in § 7 of [1] and in Toland [7]. Moreover, Theorem 5.3 of [1] states that, under some regularity conditions, the critical points of E and $\mathscr{E}$ correspond. When the corresponding critical points are nondegenerate they have the same Morse indices. In particular, they are of the same type and have the same critical values.

We shall say that the variational principle of minimizing $\mathscr{E}$ on $\mathbb{R}^n$ is a constrained variational problem if the essential domain of $\mathscr{E} = \{y \in \mathbb{R}^n : |\mathscr{E}(y)| < \infty\}$ is a proper subset of $\mathbb{R}^n$.

When E_q is given by (3.1), then the dual problem is to minimize

$$\mathscr{E}_q(y) = \frac{1}{p}\langle A^{-1}y, y\rangle^{p/2} - \frac{1}{2}\|y\|^2 \tag{9.16}$$

where $p = q' = q/(q-1)$ is the dual index to q and $1 < q < 2$.

When $q = 1$, the dual problem is to minimize

$$\mathscr{E}_1(y) = \chi_{K_2}(y) - \tfrac{1}{2}\|y\|^2 \tag{9.17}$$

where

$$K_2 = \{y \in \mathbb{R}^n : \langle A^{-1}y, y\rangle \leqq 1\}.$$

These results follow from (9.15) and the corollary to Lemma 9.2.

In particular (9.17) is a constrained variational principle but (9.16) is a new class of unconstrained variational problems for the eigenvalues and eigenvectors of A.

When F_r is given by (4.1) then the dual problem is to minimize

$$\mathscr{F}_r(y) = \chi_1(r(A)y) - \tfrac{1}{2}\|y\|^2.$$

Here χ_1 is defined as in (6.2). Equivalently we can look at

$$\tilde{\mathscr{F}}_r(z) = \chi_1(z) - \tfrac{1}{2}\|r(A)^{-1}z\|^2 \tag{9.18}$$

using the fact that $r(A)$ obeys (A3). These are constrained variational problems.

When $r(A) = (A - \mu I)$ or $(A - \mu I)^2$ as in Examples 1 or 2 of § 5, these problems are well-known inverse power methods associated with Rayleigh's principle (see Parlett [5]).

To compute the dual problem to minimizing $\mathscr{G}_p$ given by (6.1) we must first write $A = A_+ - A_-$, where A_+, A_- are defined via the spectral theorem as the parts of A

corresponding to positive and negative eigenvalues of A_-. Then $\langle Ax, x\rangle = \langle A_+x, x\rangle - \langle A_-x, x\rangle$ with both A_+, A_- being positive semidefinite. When A is positive definite, $A_- = 0$. Now

$$\mathcal{G}_p(x) = \frac{1}{p}\|x\|^p + \frac{1}{2}\langle A_-x, x\rangle - \frac{1}{2}\langle A_+x, x\rangle, \tag{9.19}$$

which has the form (9.14) with $L = I$,

$$f_1(x) = \frac{1}{p}\|x\|^p + \frac{1}{2}\langle A_-x, x\rangle,$$

$$f_2(x) = \tfrac{1}{2}\langle A_+x, x\rangle.$$

From Lemma 9.2 and (9.15), we see that the dual problem to minimizing $\mathcal{G}_p$ will be a constrained problem unless A is positive definite. When $A = A_+$ is positive definite, then the dual problem is to minimize

$$\tilde{G}_p(y) = \frac{1}{2}\langle A^{-1}y, y\rangle - \frac{1}{q}\|y\|^q \tag{9.20}$$

with $q = p/(p-1)$ being the dual index to p. This is similar to F_r defined by (4.1) with $r(A) = A^{-1}$ and with $\|x\|$ replaced by $\|y\|^q/q$ with $1 < q < 2$.

Similarly the dual principle, defined by (9.15), to minimizing G_p defined by (6.6) will be a constrained problem unless A is negative definite. When A is negative definite, the dual problem has the form (9.20) with $-A^{-1}$ in place of A^{-1}.

In summary, we have seen that the dual problems to minimizing E_q, with $1 < q < 2$ or $\mathcal{G}_p$ or G_p under definiteness constraints, are new unconstrained problems. The other dual problems are constrained variational problems, some of which are well known. Essentially the dual problem will be unconstrained if and only if both f_1^* and f_2^* in (9.15) are finite everywhere. This occurs provided both f_1 and f_2 are coercive on $\mathbb{R}^n$.

REFERENCES

[1] G. AUCHMUTY, *Duality for non-convex variational principles*, J. Differential Equations, 50 (1983), pp. 80-145.

[2] ———, *Dual variational principles for eigenvalue problems*, in Nonlinear Functional Analysis and Its Applications, Vol. I, F. E. Browder, ed., Proc. Sympos. Pure Math., 45, Part 1 (1986), pp. 55-72.

[3] ———, *Variational principles for eigenvalues of compact operators*, SIAM J. Math. Anal., 20 (1989), to appear.

[4] F. CHATELIN, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.

[5] B. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[6] G. STRANG, *Linear Algebra and Its Applications*, Academic Press, New York, 1976.

[7] J. F. TOLAND, *Duality in nonconvex optimization*, J. Math. Anal. Appl., 66 (1978), pp. 399-415.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1208–1217, September 1989

011

AN EXISTENCE AND UNIQUENESS THEOREM FOR DIFFERENCE EQUATIONS*

DARREL HANKERSON†

Abstract. The nonlinear difference equation $Py(t-k) = f(t, y(t))$ with $(j, n-j)$-conjugate boundary conditions is considered, where $Py(t-k) = 0$ is an nth-order linear difference equation and k is a fixed integer, $0 \le k < n$. Peterson considered this type of problem for the cases $j = n-1$ and $j = 1$. This paper extends his results to the $(j, n-j)$-problem. A comparison theorem for solutions of related linear inequalities is obtained, leading to some disconjugacy results. Then a shooting method type of proof is used to prove existence and uniqueness theorems for certain boundary value problems where f satisfies a two-sided Lipschitz condition.

Key words. difference equation, existence and uniqueness, disconjugate, Green's function, comparison theorem, boundary value problem

AMS(MOS) subject classifications. 34C10, 34B10, 39A10

1. Introduction.

In this paper we are interested in proving existence and uniqueness theorems for boundary value problems for the nonlinear difference equation

$$Py(t-k) = f(t, y(t)), \quad t \in [a+k, b+k], \tag{1}$$

where $f(t, y)$ is a continuous function of y for each fixed t in the interval of integers $[a+k, b+k]$. Here, k and n are fixed integers with $0 \le k < n$, and

$$Py(t) \equiv \sum_{i=0}^{n} \alpha_i(t) y(t+i), \quad t \in [a, b],$$

where $\alpha_n(t) \equiv 1$. We will be interested in $(j, n-j)$ boundary conditions of the form

$$\begin{aligned} y(a+i) &= A_i, \quad 0 \le i \le j-1, \\ y(b+n-i) &= B_i, \quad 0 \le i \le n-j-1, \end{aligned}$$

where $1 \le j \le n-1$.

A number of recent papers have considered linear and nonlinear equations of the form (1). The case $n = 2$ and $k = 1$ is discussed in [7], [8], [10], [11], and the case $n = 4$ and $k = 2$ is discussed in [3], [9], [17]. More general n and k are considered in [14]. The main existence-uniqueness theorem presented here will use a shooting method type of proof. Examples of this method applied to difference equations can be found in [1] and [14].

2. Preliminaries.

We are motivated by the results of Peterson in [14] for certain $(n-1, 1)$ and $(1, n-1)$ problems. First, we give some basic definitions. For these definitions, let L be defined by $Ly(t) = \sum_{i=0}^{n} \beta_i(t) y(t+i)$ for $t \in [a, b]$, where $\beta_n(t) \equiv 1$ and $\beta_0(t) \neq 0$ for $t \in [a, b]$.

DEFINITION (HARTMAN [5]). Let $y(t)$ be a solution of $Ly(t) = 0$. We say y has a *generalized zero* at t_0 in case either $y(t_0) = 0$ or there exists an integer j with

* Received by the editors June 8, 1987; accepted for publication (in revised form) October 31, 1988.

† Department of Algebra, Combinatorics, and Analysis, Auburn University, Auburn, Alabama 36849-5307.

$1 \le j \le t_0 - a$ such that

$$(-1)^j y(t_0 - j)y(t_0) > 0,$$
$$y(t) = 0, \quad t_0 - j < t < t_0.$$

We say that $Ly(t) = 0$ is *disconjugate* on $[a, b+n]$ if no solution $y \not\equiv 0$ has n generalized zeros.

The condition $(-1)^n \beta_0(t) > 0$ on $[a, b]$ is necessary for $Ly(t) = 0$ to be disconjugate on $[a, b + n]$. Further discussion concerning this condition can be found in [5] and [13].

DEFINITION (PETERSON [14]). Let J be a subinterval of $[a, b + n]$ with card $J \ge n$, and let $1 \le j \le n - 1$. We say that $Ly(t) = 0$ is *right* $(j, n - j)$*-disconjugate* on J provided there is no nontrivial solution $y(t)$ and integers α, $\beta \in J$ with $\alpha + j \le \beta \le \beta + n - j - 1 \in J$ such that

$$y(\alpha + i) = 0, \quad 0 \le i \le j - 1,$$
$$y(\beta + i) = 0, \quad 0 \le i \le n - j - 2,$$

and y has a generalized zero at $\beta + n - j - 1$. Similarly, we say that $Ly(t) = 0$ is *left* $(j, n - j)$*-disconjugate* on J provided there is no nontrivial solution $y(t)$ and integers α, $\beta \in J$ with $\alpha + j \le \beta \le \beta + n - j - 1 \in J$ such that

$$y(\alpha + i) = 0, \quad 0 \le i \le j - 2,$$
$$y(\beta + i) = 0, \quad 0 \le i \le n - j - 1,$$

and y has a generalized zero at $\alpha + j - 1$.

In this paper, we will say that $Ly(t) = 0$ is $(j, n - j)$*-disconjugate* on an interval J provided it is both right $(j, n - j)$-disconjugate and left $(j, n - j)$-disconjugate on J. Some results on disconjugacy and right and left disconjugacy can be found in [14] and in the references given there.

If $Ly(t) = 0$ is right $(j, n - j)$-disconjugate, then for each fixed $s \in [a, b]$, there is a unique solution $G(t, s)$ of the boundary value problem

$$\begin{aligned} &Lu(t) = \delta_{ts}, \quad t \in [a, b], \\ &u(a + i) = 0, \quad 0 \le i \le j - 1, \\ &u(b + n - i) = 0, \quad 0 \le i \le n - j - 1, \end{aligned} \tag{2}$$

where δ_{ts} is the Kronecker delta. In this case, $G(t, s)$ is called the *Green's function* for (2), and the unique solution of $Ly(t) = h(t)$ with boundary conditions in (2) is given by

$$y(t) = \sum_{s=a}^{b} G(t, s)h(s), \quad t \in [a, b + n].$$

Some of our results use the following theorem concerning the sign of the Green's function (see [15, Thm. 1]). This theorem requires that the equation $Lu(t) = 0$ be defined on a larger interval. As in [15], whenever necessary we extend the coefficients by defining $\beta_i(t) = \beta_i(a)$ for $t < a$, and $\beta_i(t) = \beta_i(b)$ for $t > b$.

THEOREM 2.1. *Assume* $2 \le j \le n - 1$ *and* $Lu(t) = 0$ *is* $(i, n - i)$*-disconjugate on* $[a + j - i, b + n + j - i]$ *for* $i = j - 1, \cdots, n - 1$. *Then the Green's function* $G(t, s)$ *for the* $(j, n - j)$*-boundary value problem*

$$\begin{aligned} &Lu(t) = \delta_{ts}, \\ &u(a + i) = 0, \quad 0 \le i \le j - 1, \\ &u(b + n - i) = 0, \quad 0 \le i \le n - j - 1, \end{aligned}$$

satisfies

$$(-1)^{n-j}G(t,s) > 0, \quad t \in [a+j, b+j],\ s \in [a,b].$$

Remark. In the case that $j = n-1$, Theorem 2.1 requires that $Lu(t) = 0$ be $(n-2,2)$-disconjugate on $[a+1, b+n+1]$. However, Peterson [12, Thm. 8] has shown that in the case of an $(n-1,1)$-problem, we can assume that the equation is $(n-2,2)$-disconjugate on $[a, b+n]$.

We will first prove some comparison theorems for solutions to certain inequalities related to the difference equations

$$Pu(t-k) = q_1(t)u(t) \tag{3}$$

$$Pv(t-k) = q_2(t)v(t) \tag{4}$$

where

$$q_1(t) \geq q_2(t), \quad t \in [a+k, b+k]. \tag{5}$$

In Peterson's paper [14], it was assumed that $1 \leq k < n$. Together with the requirement that $(-1)^n\alpha_0(t) > 0$, this guaranteed that no nontrivial solution of $Py(t-k) = p(t)y(t)$ has $n-1$ zeros at $t, \cdots, t+n-2$ and a generalized zero at $t+n-1$. In the case that $k = 0$, the corresponding condition is that $(-1)^n[\alpha_0(t) - p(t)] > 0$ for $t \in [a,b]$. Note that if $Py(t) = p(t)y(t)$ is right $(j, n-j)$-disconjugate for some j, then this condition necessarily holds.

The following is a modification of the comparison theorem in [14, Thm. 1].

THEOREM 2.2. *Assume $u(t)$ is a solution of*

$$Pu(t-k) \geq q_1(t)u(t)$$

with $u(t) \geq 0$ on $[a+k, b+k]$, and $v(t)$ is a solution of

$$Pv(t-k) \leq q_2(t)v(t)$$

such that $u(a+i) = v(a+i)$, $0 \leq i \leq n-1$. If (5) holds and (4) is right $(n-1,1)$-disconjugate on $[a, b+n]$, then $u(t) \geq v(t)$ for $t \in [a, b+n]$.

Proof. Let $w(t) = u(t) - v(t)$. Then

$$\begin{aligned} Pw(t-k) &= Pu(t-k) - Pv(t-k) \\ &\geq q_1(t)u(t) - q_2(t)v(t). \end{aligned}$$

Hence,

$$Pw(t-k) - q_2(t)w(t) \geq [q_1(t) - q_2(t)]u(t)$$

and so $w(t)$ is a solution of

$$\begin{aligned} &Pw(t-k) - q_2(t)w(t) = \Phi(t) + [q_1(t) - q_2(t)]u(t), \\ &w(a+i) = 0, \quad 0 \leq i \leq n-1, \end{aligned}$$

with $\Phi(t) \geq 0$ on $[a+k, b+k]$. By the variation of constants formula (see Peterson [14]),

$$w(t) = \sum_{s=a+k}^{t+k-1} U(t, s-k+1)[\Phi(s) + (q_1(s) - q_2(s))u(s)],$$

where $U(t,s)$ is the solution of (4) satisfying

$$U(s+i,s)=0, \quad 0\le i\le n-2,$$
$$U(s+n-1,s)=1.$$

Since (4) is right $(n-1,1)$-disconjugate, we have that $U(t,s)>0$ on $[s+n-1,b+n]$. Note that in the variation of constants formula, $t\ge s-k+1$ and all the terms are nonnegative for $s\in[a+k,b+k]$. Since the terms of the sum are understood to be zero for $s>b+k$, we obtain $w(t)\ge 0$. Hence, $u(t)\ge v(t)$ for $t\in[a,b+n]$. □

COROLLARY 2.3. *Assume* (4) *is right* $(n-1,1)$*-disconjugate on* $[a,b+n]$ *and* (5) *holds. If the coefficient* β_0 *of* $u(t-k)$ *in* (3) *satisfies* $(-1)^n\beta_0(t)>0$ *on* $[a,b]$, *then* (3) *is right* $(n-1,1)$*-disconjugate.*

The work in this paper was motivated by the following existence and uniqueness result from Peterson [14, Thm. 3]. Peterson's proof makes use of the comparison result in Theorem 2.2.

THEOREM 2.4. *Assume there is a function* $p(t)$ *defined on* $[a+k,b+k]$ *such that*

$$f(t,u)-f(t,v)\ge p(t)[u-v]$$

when $u\ge v$, $t\in[a+k,b+k]$. *If* $Py(t-k)=p(t)y(t)$ *is right* $(n-1,1)$*-disconjugate on* $[a,b+n]$, *then the boundary value problem* (1),

$$y(a+i)=A_i, \quad 0\le i\le n-2,$$
$$y(b+n)=B,$$

has a unique solution.

3. Existence and uniqueness. We wish to generalize the results of § 2 to the $(j,n-j)$-problem. We begin with the following analogue of the comparison result in Theorem 2.2.

THEOREM 3.1. *Assume* $u(t)$ *is a solution of*

$$Pu(t-k)\ge q_1(t)u(t)$$

and $v(t)$ *is a solution of*

$$Pv(t-k)\le q_2(t)v(t)$$

with $v(t)\ge 0$ *for* $t\in[a+k,b+k]$, *and*

$$u(a+i)=v(a+i), \quad 0\le i\le n-2,$$
$$u(b+n)=v(b+n).$$

If (4) *is right* $(n-1,1)$*-disconjugate on* $[a,b+n]$, (3) *is* $(n-2,2)$*-disconjugate on* $[a,b+n]$, *and* (5) *holds, then* $v(t)\ge u(t)$ *for* $t\in[a,b+n]$.

Proof. By Corollary 2.3, (3) is right $(n-1,1)$-disconjugate on $[a,b+n]$. Set $w(t)=v(t)-u(t)$ for $t\in[a,b+n]$. Then

$$Pw(t-k)=Pv(t-k)-Pu(t-k)\le q_2(t)v(t)-q_1(t)u(t)$$

and so

$$Pw(t-k)-q_1(t)w(t)\le[q_2(t)-q_1(t)]v(t).$$

Hence, $w(t)$ is a solution of

$$\begin{aligned} &Pw(t-k) - q_1(t)w(t) = \Phi(t) + [q_2(t) - q_1(t)]v(t), \\ &w(a+i) = 0, \quad 0 \le i \le n-2, \\ &w(b+n) = 0, \end{aligned} \tag{6}$$

where $\Phi(t) \le 0$ on $[a+k, b+k]$. Then

$$w(t) = \sum_{s=a+k}^{b+k} G(t, s-k)[\Phi(s) + (q_2(s) - q_1(s))v(s)]$$

where $G(t,s)$ is the Green's function for (6). Since (3) is both $(n-1,1)$-disconjugate and $(n-2,2)$-disconjugate, we have by Theorem 2.1 and Remark 2 that $G(t,s) < 0$, $t \in [a+n-1, b+n-1]$. Also, $v(t) \ge 0$ and $q_2(t) - q_1(t) \le 0$, $t \in [a+k, b+k]$, imply that $w(t) \ge 0$ for $t \in [a, b+n]$. Hence $v(t) \ge u(t)$ for $t \in [a, b+n]$. □

COROLLARY 3.2. *If* (4) *is right* $(n-1,1)$*-disconjugate and* (3) *is* $(n-2,2)$*-disconjugate and* (5) *holds, then* (4) *is right* $(n-2,2)$*-disconjugate.*

Proof. By Corollary 2.3, (3) is right $(n-1,1)$-disconjugate. Assume $c, d \in [a, b+n-1]$ with $c+n-2 < d$ and $v(t)$ is a solution of (4) with

$$\begin{aligned} &v(c+i) = 0, \quad 0 \le i \le n-3, \\ &v(c+n-2) = 1, \\ &v(d) = 0. \end{aligned}$$

It suffices to show $v(t)$ does not have a generalized zero at $d+1$. Suppose, on the contrary, that $v(t)$ has a generalized zero at $d+1$. Without loss of generality, assume d is the smallest integer greater than $c+n-2$ such that $v(d) = 0$ and v has a generalized zero at $d+1$.

First consider the case $v(t) \ge 0$ on $[c, d]$. Note that by the choice of d, $v(d-1) > 0$ and hence we must have $v(d+1) \ge 0$. Let $u(t)$ be the solution of (3) with

$$\begin{aligned} &u(c+i) = v(c+i), \quad 0 \le i \le n-2, \\ &u(d+1) = v(d+1). \end{aligned}$$

By Theorem 3.1, $v(t) \ge u(t)$ on $[c, d+1]$, and so $u(d) \le 0$. But then $u(t)$ has $n-2$ consecutive zeros followed by two generalized zeros, contradicting that (3) is both right $(n-1,1)$-disconjugate and right $(n-2,2)$-disconjugate (see [12, Thm. 7]).

Now consider the case $v(t) < 0$ for some $t \in [c+n-1, d-1]$. Let $v_1(t)$ be the solution of (4),

$$\begin{aligned} &v(c+i) = 0, \quad 0 \le i \le n-2, \\ &v(c+n-1) = -1. \end{aligned}$$

Then $v_1(t) < 0$ on $[c+n-1, d]$ and there exists $\alpha > 0$ such that

$$w(t) \equiv v(t) - \alpha v_1(t) \ge 0, \quad t \in [c, d]$$

with equality holding at some $t_0 \in [c+n-1, d-1]$. We can assume $t_0 \ge c+n-1$ is the first point such that $w(t_0) = 0$. Then $w(t) \ge 0$ satisfies

$$\begin{aligned} &Pw(t-k) = q_2(t)w(t), \\ &w(c+i) = 0, \quad 0 \le i \le n-3, \\ &w(c+n-2) = 1, \\ &w(t_0) = 0, \end{aligned}$$

and $w(t)$ has a generalized zero at t_0+1. Applying the first case to $w(t)$ on $[c, t_0+1]$ leads to a contradiction. □

In some of the proofs which follow, uniqueness of solutions to certain boundary value problems for (1) is coupled with the Brouwer Invariance of Domain Theorem [2] to obtain that solutions depend continously on boundary conditions. A typical argument verifying this continuous dependence on boundary conditions can be found in [6].

We are now ready to prove an existence-uniqueness theorem for the $(n-2,2)$-problem. This theorem is an extension of Theorem 2.4.

THEOREM 3.3. *Assume there exist functions $p(t)$ and $r(t)$ defined on $[a+k, b+k]$ such that*

$$p(t)[u-v] \le f(t,u) - f(t,v) \le r(t)[u-v]$$

whenever $u \ge v$, $t \in [a+k, b+k]$. If $Py(t-k) = p(t)y(t)$ is right $(n-1,1)$-disconjugate on $[a, b+n]$ and $Pu(t-k) = r(t)u(t)$ is $(n-2,2)$-disconjugate on $[a, b+n]$, then the boundary value problem (1),

$$\begin{aligned} &y(a+i) = A_i, \quad 0 \le i \le n-3, \\ &y(b+n-i) = B_i, \quad 0 \le i \le 1, \end{aligned} \tag{7}$$

has a unique solution.

Proof. We will use the shooting method in this proof. By Theorem 2.4, there exists a unique solution $y(t,m)$ of the boundary value problem (1),

$$\begin{aligned} &y(a+i) = A_i, \quad 0 \le i \le n-3, \\ &y(a+n-2) = m, \\ &y(b+n) = B_0. \end{aligned}$$

Let $S = \{y(b+n-1, m) \mid m \in \mathbf{R}\}$. A standard argument using the Brouwer Invariance of Domain Theorem shows that $y(b+n-1, m)$ is a continuous function of m. Hence, S is a nonempty connected set. Thus to prove existence, it suffices to show that S is neither bounded below nor above.

Fix $m_1 > m_2$ and set

$$z(t) = \frac{y(t,m_1) - y(t,m_2)}{m_1 - m_2}.$$

Note that

$$\begin{aligned} Pz(t-k) &= \frac{Py(t-k,m_1) - Py(t-k,m_2)}{m_1 - m_2} \\ &= \frac{f(t, y(t,m_1)) - f(t, y(t,m_2))}{m_1 - m_2}. \end{aligned}$$

Now define

$$q(t) = \begin{cases} \dfrac{f(t, y(t,m_1)) - f(t, y(t,m_2))}{y(t,m_1) - y(t,m_2)}, & y(t,m_1) \ne y(t,m_2), \\ p(t), & y(t,m_1) = y(t,m_2). \end{cases}$$

Then $z(t)$ is the solution of the boundary value problem

$$\begin{aligned} &Pz(t-k) = q(t)z(t), \\ &z(a+i) = 0, \quad 0 \le i \le n-3, \\ &z(a+n-2) = 1, \\ &z(b+n) = 0. \end{aligned}$$

Note also that $p(t) \leq q(t) \leq r(t)$ for $t \in [a+k, b+k]$. Since $Py(t-k) = p(t)y(t)$ is right $(n-1,1)$-disconjugate, we have that both

$$Pz(t-k) = q(t)z(t) \tag{8}$$

and

$$Pu(t-k) = r(t)u(t) \tag{9}$$

are right $(n-1,1)$-disconjugate by Corollary 2.3. Now use the fact that (9) is also $(n-2,2)$-disconjugate and apply Corollary 3.2 to obtain that (8) is right $(n-2,2)$-disconjugate. It follows that $z(t) \geq 0$ on $[a, b+n]$. Next, apply Theorem 3.1 to obtain that $z(t) \geq u(t)$, where $u(t)$ is the solution of (9),

$$\begin{aligned} &u(a+i) = z(a+i), \quad 0 \leq i \leq n-2, \\ &u(b+n) = z(b+n). \end{aligned}$$

Hence,

$$y(t, m_1) - y(t, m_2) \geq u(t)(m_1 - m_2)$$

for $t \in [a, b+n]$. Letting $t = b+n-1$ we obtain

$$y(b+n-1, m_1) - y(b+n-1, m_2) \geq u(b+n-1)(m_1 - m_2). \tag{10}$$

Since (9) is both $(n-1,1)$ and $(n-2,2)$-disconjugate on $[a, b+n]$, it follows from [12, Thm. 7] that $u(t) > 0$ on $[a+n-2, b+n-1]$. Letting $m_1 \to \infty$ we see that

$$\lim_{m \to \infty} y(b+n-1, m) = \infty,$$

and similarly, letting $m_2 \to -\infty$ we see that

$$\lim_{m \to -\infty} y(b+n-1, m) = -\infty.$$

It follows that $S = \mathbf{R}$ and the existence part of the proof is complete.

For the uniqueness of solutions, suppose on the contrary that $y_1(t)$ and $y_2(t)$ are distinct solutions of the boundary value problem (1), (7). Since solutions to the $(n-1,1)$-problem are unique, we can write $y_1(t) = y(t, m_1)$, $y_2(t) = y(t, m_2)$, for some $m_1 \neq m_2 \in \mathbf{R}$. Without loss of generality, we can assume that $m_1 > m_2$. But then (10) shows that $y(b+n-1, m_1) \neq y(b+n-1, m_2)$, contradicting the assumption that both $y_1(t)$ and $y_2(t)$ were solutions to the same $(n-2,2)$-problem. Hence the uniqueness condition is satisfied and the proof is complete. □

Example. Let Δ be defined by $\Delta y(t) = y(t+1) - y(t)$ and consider the difference equation $\Delta^5 y(t-1) = e^{1-t} \sin y(t)$ for $t \in [1,4]$. Note that for $u \geq v$, $-[u-v] \leq e^{1-t}(\sin u - \sin v) \leq [u-v]$ for $t \in [1,4]$. Using [16, Thm. 2], we can show that the equation $\Delta^5 y(t-1) = -y(t)$ is right $(4,1)$-disconjugate on $[0,8]$, and $\Delta^5 y(t-1) = y(t)$ is $(3,2)$-disconjugate on $[0,8]$. By Theorem 3.3, there is a unique solution of the boundary value problem

$$\begin{aligned} &\Delta^5 y(t-1) = e^{1-t} \sin y(t), \quad t \in [1,4], \\ &y(i) = A_i, \quad 0 \leq i \leq 2, \\ &y(8-i) = B_i, \quad 0 \leq i \leq 1. \end{aligned}$$

We conclude with a comparison theorem and an existence-uniqueness theorem for the general case.

THEOREM 3.4. *Let* $1 \le j \le n-1$. *Assume* $u(t)$ *is a solution of*

$$Pu(t-k) \ge q_1(t)u(t)$$

and $v(t)$ *is a solution of*

$$Pv(t-k) \le q_2(t)v(t)$$

with

$$\begin{aligned} &u(a+i) = v(a+i), \quad 0 \le i \le j-1, \\ &u(b+n-i) = v(b+n-i), \quad 0 \le i \le n-j-1. \end{aligned}$$

If one of the following holds:

1. $u(t) \ge 0$ *on* $[a+k, b+k]$ *and* (4) *is either disconjugate on* $[a, b+n]$ *or* $(i, n-i)$*-disconjugate on* $[a+j-i, b+n+j-i]$ *for* $j-1 \le i < n$, *or*
2. $v(t) \ge 0$ *on* $[a+k, b+k]$ *and* (3) *is either disconjugate on* $[a, b+n]$ *or* $(i, n-i)$*-disconjugate on* $[a+j-i, b+n+j-i]$ *for* $j-1 \le i < n$,

then $(-1)^{n-j}u(t) \ge (-1)^{n-j}v(t)$ *for* $t \in [a, b+n]$.

Proof. First assume that condition 1 holds. Set $w(t) = u(t) - v(t)$. Then

$$Pw(t-k) = Pu(t-k) - Pv(t-k) \ge q_1(t)u(t) - q_2(t)v(t)$$

and so

$$Pw(t-k) - q_2(t)w(t) \ge (q_1(t) - q_2(t))u(t).$$

Hence, $w(t)$ is a solution of

$$\begin{aligned} &Pw(t-k) - q_2(t)w(t) = \Phi(t) + [q_1(t) - q_2(t)]u(t), \\ &w(a+i) = 0, \quad 0 \le i \le j-1, \\ &w(b+n-i) = 0, \quad 0 \le i \le n-j-1, \end{aligned} \tag{11}$$

where $\Phi(t) \ge 0$ on $[a+k, b+k]$. Then

$$w(t) = \sum_{s=a+k}^{b+k} G(t, s-k)[\Phi(s) + (q_1(s) - q_2(s))u(s)]$$

where $G(t,s)$ is the Green's function for (11). By the disconjugacy assumptions on (4), we have that $(-1)^{n-j}G(t,s) > 0$, $t \in [a+j, b+j]$. Since $u(t) \ge 0$ and $q_1(t) - q_2(t) \ge 0$ for $t \in [a+k, b+k]$, it follows that $(-1)^{n-j}w(t) \ge 0$ for $t \in [a, b+n]$. Hence, in this case, $(-1)^{n-j}u(t) \ge (-1)^{n-j}v(t)$ for $t \in [a, b+n]$.

The proof in the case where condition 2 holds is similar and will be omitted. □

THEOREM 3.5. *Let* $1 \le j \le n-1$. *Assume there exist functions* $p(t)$, $r(t)$, *defined on* $[a,b]$ *such that*

$$p(t)[u-v] \le f(t,u) - f(t,v) \le r(t)[u-v]$$

whenever $u \ge v$, $t \in [a,b]$. *If* $Py(t) = p(t)y(t)$ *and* $Py(t) = r(t)y(t)$ *are disconjugate on* $[a, b+n]$, *then the boundary value problem*

$$\begin{aligned} &Py(t) = f(t, y(t)), \quad t \in [a,b], \\ &y(a+i) = A_i, \quad 0 \le i \le j-1, \\ &y(b+n-i) = B_i, \quad 0 \le i \le n-j-1, \end{aligned}$$

has a unique solution.

Proof. The proof is by induction on decreasing values of j. The cases $j = n-1$ and $j = n-2$ are contained in Theorems 2.4 and 3.3. Assume $j \leq n-3$ and that the theorem holds if j is replaced by $j+1$. Then there exists a unique solution $y(t,m)$ to the $(j+1, n-j-1)$-problem

$$\begin{aligned} &Py(t) = f(t, y(t)), \quad t \in [a,b], \\ &y(a+i) = A_i, \quad 0 \leq i \leq j-1, \\ &y(a+j) = m, \\ &y(b+n-i) = B_i, \quad 0 \leq i \leq n-j-2. \end{aligned}$$

Fix $m_1 > m_2$ and define $z(t)$ and $q(t)$ as in the proof of Theorem 3.3. Then $z(t)$ is the solution of the boundary value problem

$$\begin{aligned} &Pz(t) = q(t)z(t), \\ &z(a+i) = 0, \quad 0 \leq i \leq j-1, \\ &z(a+j) = 1, \\ &z(b+n-i) = 0, \quad 0 \leq i \leq n-j-2. \end{aligned}$$

Note that $p(t) \leq q(t) \leq r(t)$ for $t \in [a,b]$. By results of Eloe [4] and Peterson [13], the equation $Pz(t) = q(t)z(t)$ is disconjugate on $[a, b+n]$.

To simplify the notation, we consider the case that $n-j$ is odd; the case for $n-j$ even is similar. Let $v(t)$ be the solution of

$$\begin{aligned} &Pv(t) = p(t)v(t), \\ &v(a+i) = z(a+i), \quad 0 \leq i \leq j, \\ &v(b+n-i) = z(b+n-i), \quad 0 \leq i \leq n-j-2. \end{aligned}$$

By the disconjugacy assumptions, we have $v(t) \geq 0$ for $t \in [a, b+n]$. By Theorem 3.4, it follows that $z(t) \geq v(t)$ for $t \in [a, b+n]$.

The remainder of the proof can be modeled after the proof of Theorem 3.3. □

REFERENCES

[1] R. P. AGARWAL, *Computational methods for discrete boundary value problems*, Appl. Math. Comput., 18 (1986), pp. 15–41.

[2] E. ARTIN AND H. BRAUN, *Introduction to Algebraic Topology*, Merrill, Columbus, Ohio, 1969.

[3] S. CHENG, *On a class of fourth order linear recurrence equations*, Internat. J. Math. Math. Sci., 7 (1984), pp. 131–149.

[4] P. W. ELOE, *A comparison theorem for linear difference equations*, Proc. Amer. Math. Soc., 103 (1988), pp. 451–457.

[5] P. HARTMAN, *Difference equations: disconjugacy, principal solutions, Green's functions, complete monotonicity*, Trans. Amer. Math. Soc., 246 (1978), pp. 1–30.

[6] J. HENDERSON AND L. JACKSON, *Existence and uniqueness of solutions of k-point boundary value problems for ordinary differential equations*, J. Differential Equations, 48 (1983), pp. 373–385.

[7] D. B. HINTON AND R. T. LEWIS, *Spectral analysis of second order difference equations*, J. Math. Anal. Appl., 63 (1978), pp. 421–438.

[8] J. W. HOOKER, *A Hille–Wintner type comparison theorem for second order difference equations*, Internat. J. Math. Math. Sci., 6 (1983), pp. 387–394.

[9] J. W. HOOKER AND W. T. PATULA, *Growth and oscillation properties of solutions of a fourth order linear difference equation*, J. Austral. Math. Soc. Ser. B, 26 (1985), pp. 310–328.

[10] ———, *Riccati type transformations for second-order linear difference equations*, J. Math. Anal. Appl., 82 (1981), pp. 451–462.

[11] J. W. HOOKER AND W. T. PATULA, *A second-order nonlinear difference equation: oscillation and asymptotic behavior*, J. Math. Anal. Appl., 91 (1983), pp. 9–29.

[12] A. PETERSON, *Boundary value problems and Green's functions for linear difference equations*, in Differential and Integral Equations, Proceedings of the Twelfth and Thirteenth Midwest Conferences, J. L. Henderson, ed., 1985, pp. 79–100.

[13] ———, *A comparison theorem for linear difference equations*, in Proceedings of the International Symposium on Nonlinear Analysis and Applications to Biomathematics, Andhra University, Visakhapatnam, India, 1987.

[14] ———, *Existence and uniqueness theorems for nonlinear difference equations*, J. Math. Anal. Appl., 125 (1987), pp. 185–191.

[15] ———, *Green's functions for $(k, n-k)$-boundary value problems for linear difference equations*, J. Math. Anal. Appl., 124 (1987), pp. 127–138.

[16] ———, *On $(k, n-k)$-disconjugacy for linear difference equations*, in Qualitative Properties of Differential Equations, Proceedings of the 1984 Edmonton Conference, W. Allegretto and G. J. Butler, eds., 1986, pp. 329–337.

[17] B. SMITH AND W. E. TAYLOR, JR., *Oscillatory and asymptotic behavior of certain fourth order difference equations*, Rocky Mountain J. Math., 16 (1986), pp. 403–406.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1218-1242, September 1989

012

HIDDEN VARIABLE FRACTAL INTERPOLATION FUNCTIONS*

M. F. BARNSLEY†, J. ELTON†, D. HARDIN‡, AND P. MASSOPUST§

Abstract. Interpolation functions $f: [0, 1] \to \mathbb{R}$ of the following nature are constructed. Given data

$$\{(t_n, x_n) \in [0, 1] \times \mathbb{R}: n = 0, 1, 2, \cdots, N\}$$

with $0 = t_0 < t_1 < \cdots < t_N = 1$, f obeys

$$f(t_n) = x_n, \qquad n = 0, 1, 2, \cdots, N.$$

Furthermore, the graph of f is the projection of a set G in $\mathbb{R}^M$ (M an integer greater than or equal to 2) that is homeomorphic to [0, 1] and is the attractor for an iterated function system consisting of affine maps in $\mathbb{R}^M$. The latter characterization ensures that f can be computed rapidly while possessing many "hidden" variables, on which its values continuously depend, which allow great flexibility and diversity in the interpolant, making it potentially useful in approximation theory. Estimates and exact values for the fractal dimensions of G and the graph of f are obtained.

Key words. fractal, self-affine, iterated function systems, fractal dimension, fractal interpolation

AMS(MOS) subject classifications. 26A30, 41A30, 58F12, 58F13

1. Introduction.

A function $F: [0, 1] \to \mathbb{R}^n$ is self-affine if its graph G obeys

$$G = \cup\, w_i(G)$$

where the union is over a finite set of affine maps $w_i: \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$. There has been recent interest in such functions with $n = 1$ because of their potential utility in approximation theory [B], [BH] and in computer graphics [DHN]. Specific properties when $n = 1$ include the availability of fast algorithms for computing the values of such functions; the possibility of choosing the w_i so that F interpolates given values; and the computability of their moments $\int_0^1 x^n F(x)\, dx$, their indefinite integrals

$$\int_0^x \int_0^{x_1} \cdots \int_0^{x_p} F(x_{p+1})\, dx_{p+1}\, dx_p \cdots dx_1,$$

and their Fourier transforms, $\hat{F}(k) = \int_0^1 e^{ikx} F(x)\, dx$. Moreover, in the case of equally spaced interpolation points, the fractal dimension of these functions can be expressed in terms of the coefficients in the affine transformations. Such functions yield solutions to special types of functional equations [H], and have been related to special types of dynamical systems [HM] for which Lyapunov exponent calculations can be made.

Here our main aim is to show how the class of one-dimensional interpolation functions can be usefully widened by considering the projections of the graphs of higher-dimensional self-affine functions. We show that we still have the advantages of rapid computability, the feasibility of interpolation, and the ability to express the fractal dimension in terms of underlying parameters. The class of functions obtained is much more diverse because their values depend continuously on all of the "hidden" variables, namely, the coefficients of the possibly high-dimensional affine maps that determine the function. In pursuing our main aim we will reveal a number of facts about multidimensional curves, including space-filling and fractal-filling curves, that are surprising and suggest new lines of investigation and applications.

* Received by the editors October 20, 1986; accepted for publication (in revised form) November 18, 1988.
† School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.
‡ Department of Mathematics, Vanderbilt University, Nashville, Tennessee 37235.
§ Department of Mathematics, La Grange College, La Grange, Georgia 30240.

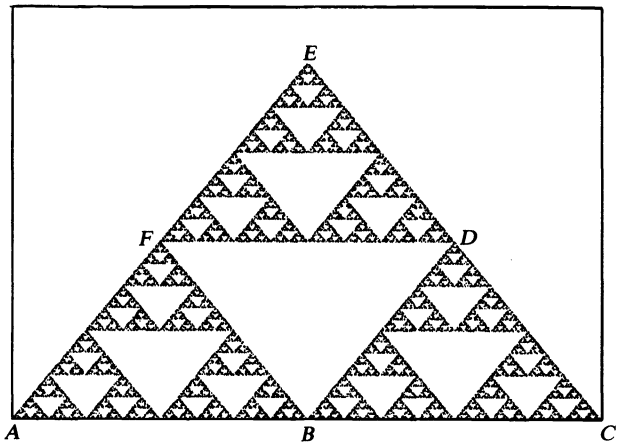


FIG. 1. *The Sierpinski triangle. A, B, C, D, E, F label vertices of three subtriangles.*

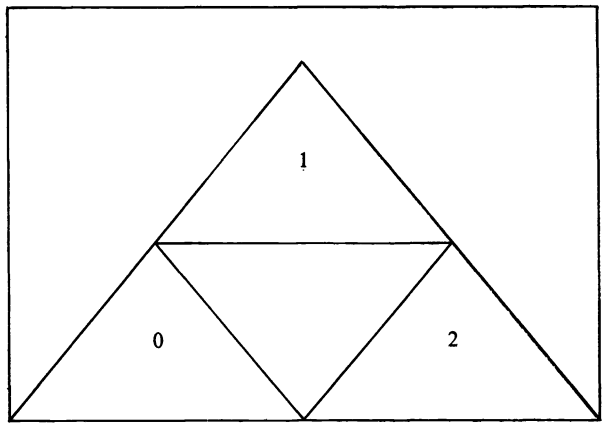


FIG. 2. 0, 1, *and* 2 *label the images of the large triangle under maps* w_0, w_1, *and* w_2, *respectively.*

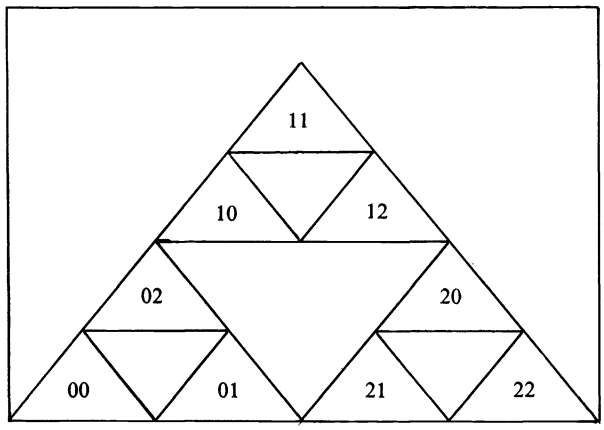


FIG. 3. 00, 01, $\cdots$, 22 *label the images of the triangles in Fig.* 2 *under the maps* w_0, w_1, *and* w_2.

We show how the class of functions introduced here was conceived. Let $S \subset \mathbb{R}^2$ denote the Sierpinski triangle, illustrated in Fig. 1 and specified more precisely below. Label points A, B, C, D, E, and F on S as in Fig. 1. Let w_0, w_1, w_2 be three affine transformations $w_i: \mathbb{R}^2 \to \mathbb{R}^2$, of the form

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

where a, b, c, d, e, and f are real constants. The maps are uniquely specified by the requirements $w_0(A) = A$, $w_0(E) = B$, $w_0(C) = F$, $w_1(A) = F$, $w_1(E) = E$, $w_1(C) = D$, $w_2(A) = D$, $w_2(E) = B$, $w_2(C) = C$. Thus w_0 takes the triangle ACE onto the subtriangle AFB, w_1 takes ACE onto FDE, and w_2 takes ACE onto BDC. The orientations are important as we will see. Each map is a strict contraction, and S is the unique compact subset of $\mathbb{R}^2$ so that [BD], [H]

$$S = \bigcup_{i=0}^{2} w_i(S).$$

We now construct a continuous map ϕ of $[0, 1]$ onto S. Each point of S possesses at least one address, which consists of a triadic decimal $\sigma = .\sigma_1\sigma_2\sigma_3 \cdots \sigma_n \cdots$, where $\sigma_i \in \{0, 1, 2\}$, and each $\sigma \in [0, 1]$ corresponds to a unique point on S (see [BD], [H]). The point $\phi(\sigma) \in S$ corresponding to $\sigma \in [0, 1]$ may be computed by

$$\phi(\sigma) = \lim_{n \to \infty} w_{\sigma_1} \circ w_{\sigma_2} \circ \cdots \circ w_{\sigma_n}(A).$$

This addressing system is readily understood: the original large triangle is mapped onto three smaller triangles labeled 0, 1, and 2 under maps w_0, w_1, and w_2, respectively; see Fig. 2. All addresses that begin with .0 correspond to points in region 0. Now apply all three maps to the latter subtriangles to obtain nine smaller triangles labeled 00, 01, $\cdots$, 22 as illustrated in Fig. 3. All addresses that begin with .12 correspond to points that lie in the region labeled 12.

We can show [H], [B] that the map $\phi: [0, 1] \to S$ is continuous; see also § 2.2. The curve obtained can be thought of as a "space"-filling curve where the space is S with fractal dimension $\log 3/\log 2$. This curve can also be obtained as the limit of the recursive refinement procedure suggested in Fig. 4; we find a sequence of piecewise linear functions $\phi_n: [0, 1] \to S$ that converges uniformly to ϕ.

Now let us write $\phi(t) = (f(t), h(t))$. Then $f: [0, 1] \to \mathbb{R}$ is an example of the type of function discussed in this paper, which we call a hidden variable fractal interpolation function. A plot of $f(t)$ corresponding to $A = (0, 0)$, $C = (1, 0)$, and $E = (\frac{1}{2}, \sqrt{3}/2)$ is shown in Fig. 5. First, we note that $f(t)$ is not self-affine: its graph does not consist of the union of affine images of itself. Second, $f(t)$ can be computed using standard iterated function system (IFS) algorithms, which are rapid because the graph of f is the projection of the attractor for a set of three-dimensional affine maps. Third, the fractal dimension of f can be computed exactly (it is $2 - \log_3 2$). Fourth, given any set of data (t_i, x_i) for $i = 0, 1, 2, \cdots, N$, a generalization of the above procedure yields a function f that interpolates the data while retaining many degrees of freedom.

The structure of this paper is as follows. In § 2 we review iterated function system (IFS) theory, including the basic characterization and computation of attractors of IFS; then we give the basic theorem on the existence of attractors of IFS that are graphs of certain functions with domains and ranges in compact metric spaces. These domains and ranges may themselves be fractal objects, characterized as attractors of IFS. We show, in particular, the construction by which these graphs can be made to

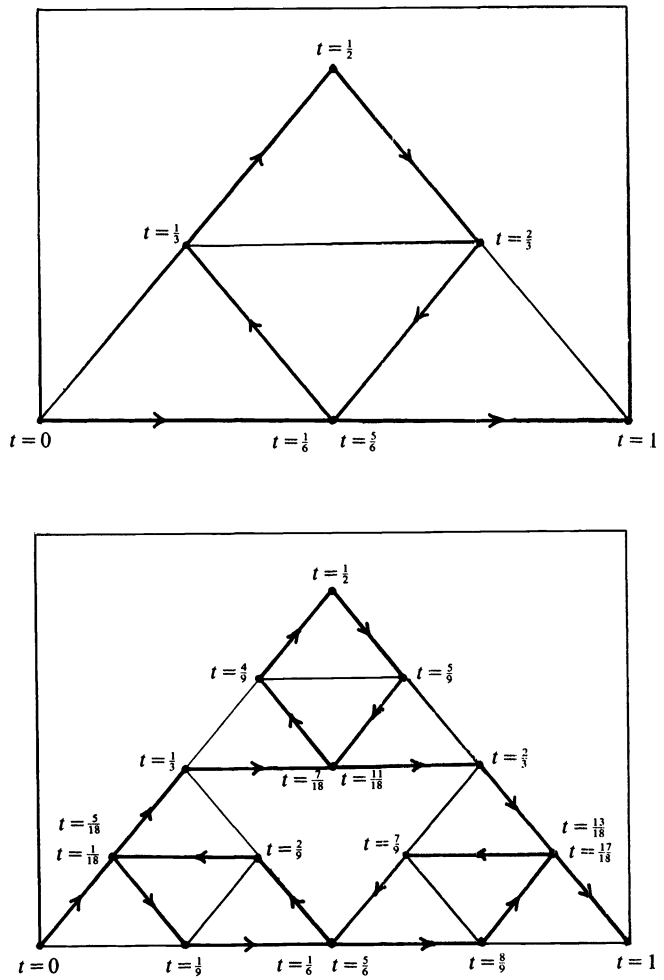


FIG. 4. *Illustration of the generation of the "space-filling" curve* $\phi: [0, 1] \to S$ *by recursive refinement.*

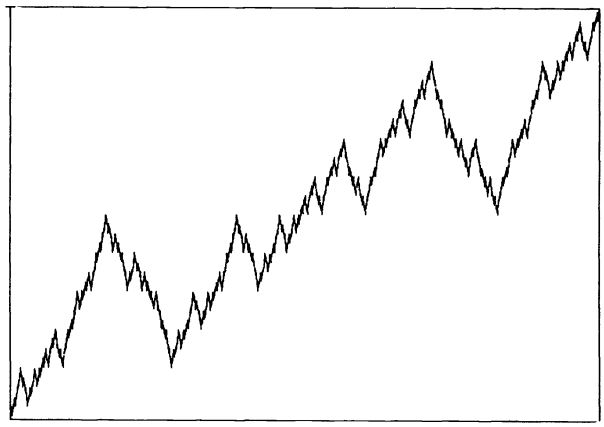


FIG. 5. *Plot of a hidden variable fractal function corresponding to a Sierpinski triangle, as described in the text.*

interpolate data. Examples are given to illustrate the appearance and diversity of these functions.

In § 3 we estimate the fractal dimension of the graphs of certain functions that map attractors of IFS onto attractors of IFS, in the case where the IFSs are formed of similitudes that contract more on one attractor than on the other. When a certain open set condition prevails the estimate is the actual value.

Section 4 is devoted to the main hard theorem: (a) we prove the conjecture of Hardin and Massopust [HM] on the fractal dimension of self-affine functions in one dimension; (b) we compute the fractal dimension of certain hidden variable interpolation functions.

2. Construction of fractal interpolation functions.

2.1. Attractors of IFS and the deterministic algorithm for their evaluation. Let K be a compact metric space with metric $d(\cdot, \cdot)$. Let $w_i: K \to K$ for $i = 1, 2, \cdots, N$, where N is a finite positive integer, be strict contractions; that is, there exists a constant $0 \leqq s < 1$ such that

$$d(w_i(x), w_i(y)) \leqq s \cdot d(x, y) \quad \forall x, y \in K, \quad \forall i.$$

We call $(K, w_i: i = 1, 2, \cdots, N)$ an IFS. (This definition of an iterated function system is more restrictive than elsewhere (see [BD], for example) but serves well for this paper.) There exists a unique compact set $A \subset K$ such that

$$A = \cup\, w_i(A).$$

A is called the attractor of the IFS.

Let H denote the set of nonempty compact subsets of K, and endow it with the Hausdorff metric

$$h(B, C) = \max\left\{\max_{x \in B} \min_{y \in C} d(x, y), \max_{y \in C} \min_{x \in B} d(x, y)\right\}$$

for all $B, C \in H$. Then H is a complete metric space. Define $W: H \to H$ by

$$W(B) = \cup\, w_i(B).$$

Then W is a strict contraction with

$$h(W(B), W(C)) \leqq s \cdot h(B, C) \quad \forall B, C \in H;$$

and A is the unique fixed point of W.

The deterministic algorithm for the computation of A is to choose any $A_0 \in H$ and define

$$A_{n+1} = W(A_n).$$

Then

$$A = \lim_{n \to \infty} A_n.$$

Suppose, for example, that $K = [0, 1] \times [0, 1] \subset \mathbb{R}^2$. Then $h(A_n, A) \leqq s^n(1 - s)$, whence the number of iterations needed to evaluate A to a desired precision is readily calculated.

The deterministic algorithm is well suited to the evaluation of A when it is the graph of a function in $\mathbb{R}^n$. An illustration is given in Fig. 6, where we show the computation of an attractor for two affine maps in $\mathbb{R}^2$. In other situations, for example, when A represents a digitized image, or not all the maps are contractions, or we are concerned with measures supported on A, the random iteration algorithm (described, for example, in [BD]) is more appropriate.

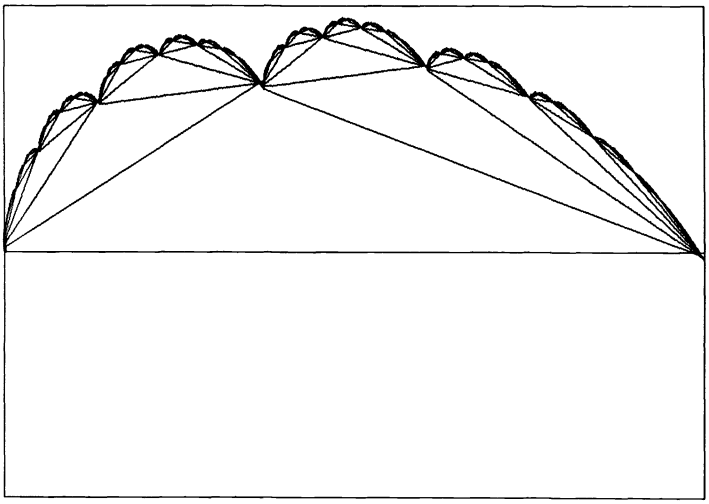


FIG. 6. *A sequence of graphs of functions* $A_0, A_1, A_2, \cdots$, *computed using the deterministic algorithm, that converge to the attractor* A *of an* IFS *in* $\mathbb{R}^2$ *consisting of a pair of affine maps.*

2.2. Construction of functions using IFS. In this section we present a general construction for an IFS whose attractor is the graph of a function $F: I \to K$ where J $(\supset I)$ and K are compact metric spaces.

We use $d(\cdot, \cdot)$ for the distance function in both J and K; the argument will specify which function is applied. We define a distance

$$d((t_1, X_1), (t_2, X_2)) = d(t_1, t_2) + \theta d(X_1, X_2) \tag{2.1}$$

between points (t_1, X_1) and (t_2, X_2) in $J \times K$, where $\theta > 0$ is to be specified. Then $J \times K$ is a compact metric space.

Let $l: J \to J$ be a strict contraction: there is a constant $0 \leqq s_1 < 1$ so that

$$d(l(t_1), l(t_2)) \leqq s_1 \cdot d(t_1, t_2) \quad \forall t_1, t_2 \in J. \tag{2.2}$$

Let $k: J \times K \to K$ and let there be constants c and s_2, with $0 \leqq s_2 < 1$, so that

$$d(k(t_1, X), k(t_2, X)) \leqq c \cdot d(t_1, t_2) \quad \forall t_1, t_2 \in J, \quad X \in K, \tag{2.3}$$

$$d(k(t, X_1), k(t, X_2)) \leqq s_2 \cdot d(X_1, X_2) \quad \forall t \in J, \quad X_1, X_2 \in K. \tag{2.4}$$

LEMMA 2.1. *If* θ *in* (2.1) *is equal to* $(1 - s_1)/2c$, *then* $W: J \times K \to J \times K$ *defined by* $W(t, X) = (l(t), k(t, X))$, *with* l *and* k *as above, is a strict contraction.*

Proof.

$$\begin{aligned} d(W(t_1, X_1), W(t_2, X_2)) &= d(l(t_1), l(t_2)) + \theta d(k(t_1, X_1), k(t_2, X_2)) \\ &\leqq s_1 d(t_1, t_2) + \theta d(k(t_1, X_1), k(t_2, X_1)) \\ &\qquad + \theta d(k(t_2, X_1), k(t_2, X_2)) \\ &\leqq (s_1 + \theta c) d(t_1, t_2) + \theta s_2 d(X_1, X_2) \\ &\leqq q(d(t_1, t_2) + \theta d(X_1, X_2)) \\ &= q d((t_1, X_1), (t_2, X_2)) \end{aligned}$$

where $q = \max\{(s_1 + \theta c), s_2\} < 1$. □

Now we consider the following construction. Let $W_i: J \times K \to J \times K$ for $i = 1, 2, \cdots, N$ be a finite set of maps of the structure

$$W_i(t, X) = (l_i(t), k_i(t, X)) \tag{2.5}$$

where l_i, k_i obey (2.2)-(2.4), and let θ in (2.1) be chosen according to Lemma 2.1. Then $(J \times K, W_i: i = 1, 2, \cdots, N)$ is an IFS and possesses a unique attractor G. Also $(J, l_i: i = 1, 2, \cdots, N)$ is an IFS; we denote its attractor by I. Let $P_t: J \times K \to J$ be the projection operator defined by $P_t(t, X) = t$ for all $(t, X) \in J \times K$. Then clearly $P_t G = I$. We will give conditions so that G is the graph of a function $F: I \to K$.

Let

$$(t_j, X_j) \in J \times K \quad \text{for } j = 0, 1, 2, \cdots, N \tag{2.6}$$

and suppose

$$W_i(t_0, X_0) = (t_{i-1}, X_{i-1}) \quad \text{and} \quad W_i(t_N, X_N) = (t_i, X_i) \quad \text{for } i = 1, 2, \cdots, N. \tag{2.7}$$

Then $\{(t_j, X_j): i = 0, 1, \cdots, N\} \subset G$ because $(t_0, X_0) \in G$ as it is the fixed point of W_1, $(t_N, X_N) \in G$ as it is the fixed point of W_N, and $(t_i, X_i) \in G$ for $i = 1, 2, \cdots, N-1$ as $W_i(t_N, X_N) = (t_i, X_i)$.

Now assume that the maps $l_i: I \to I$ are invertible over their ranges $l_i(I)$, that

$$l_i(I) \cap l_j(I) = \phi \quad \text{when } |i - j| \notin \{0, 1\}, \tag{2.8}$$

$$l_i(I) \cap l_{i+1}(I) = t_i \quad \text{for } i = 1, 2, \cdots, N-1. \tag{2.9}$$

THEOREM 1. *Let $\{J \times K, W_i: i = 1, 2, \cdots, N\}$ be an* IFS *of the special structure described above, so that it obeys* (2.5)-(2.9). *Then its attractor G is the graph of a continuous function $F: I \to K$ such that*

$$F(t_j) = X_j \quad \textit{for } j = 0, 1, 2, \cdots, N.$$

Proof. I is a compact metric space with the distance function d inherited from J. Let $\mathcal{F} = \{f: I \to K: f \text{ is continuous on } I, f(t_0) = X_0, f(t_N) = X_N\}$. Define a metric d on $\mathcal{F}$ by

$$d(f_1, f_2) = \max\{d(f_1(t), f_2(t)): t \in I\}.$$

Then $\mathcal{F}$ is a complete metric space because the uniform limit of continuous functions is continuous. Define $T: \mathcal{F} \to \mathcal{F}$ by

$$(Tf)(t) = k_i(l_i^{-1}(t), f(l_i^{-1}(t))) \quad \text{for } t \in l_i(I), f \in \mathcal{F}.$$

(Recall that $I = \cup\, l_i(I)$.) We show that T is well defined and that it takes $\mathcal{F} \to \mathcal{F}$. In view of (2.8) and (2.9), $(Tf)(t)$ is well defined at points $t \in I \backslash \{t_0, t_1, \cdots, t_N\}$. Also, it is continuous on each of the compact sets $l_i(t)$. Now observe that

$$\begin{aligned}
\lim_{\substack{t \to t_i \\ t \in l_i(I)}} (Tf)(t) &= k_i(l_i^{-1}(t_i), f(l_i^{-1}(t_i))) \\
&= k_i(t_N, f(t_N)) = k_i(t_N, X_N) \\
&= k_{i+1}(t_0, x_0) = k_{i+1}(t_0, f(t_0)) \\
&= k_{i+1}(l_{i+1}^{-1}(t_i), f(l_{i+1}^{-1}(t_i))) = \lim_{\substack{t \to t_i \\ t \in l_{i+1}(I)}} (Tf)(t),
\end{aligned}$$

which shows $(Tf)(t)$ is both well defined and continuous at each of the points $\{t_0, t_1, \cdots, t_N\}$. Moreover,

$$\begin{aligned}(Tf)(t_0) &= k_1(l_1^{-1}(t_0), f(l_1^{-1}(t_0))) \\ &= k_1(t_0, f(t_0)) = k_1(t_0, X_0) = X_0, \\ (Tf)(t_N) &= k_N(l_N^{-1}(t_N), f(l_N^{-1}(t_N))) \\ &= k_N(t_N, f(t_N)) \\ &= k_N(t_N, X_N) = X_N,\end{aligned}$$

so indeed $Tf \in \mathcal{F}$ when $f \in \mathcal{F}$.

We now show that T is a strict contraction in $\mathcal{F}$. For $t \in l_i(I)$ and $f_1, f_2 \in \mathcal{F}$, we have

$$\begin{aligned}d((Tf_1)(t), (Tf_2)(t)) &= d(k_i(l_i^{-1}(t), f_1(l_i^{-1}(t))), k_i(l_i^{-1}(t), f_2(l_i^{-1}(t)))) \\ &\leqq s_2 \cdot d(f_1(l_i^{-1}(t)), f_2(l_i^{-1}(t))) \\ &\leqq s_2 \cdot d(f_1, f_2)\end{aligned}$$

whence $d(Tf_1, Tf_2) \leqq s_2 d(f_1, f_2)$. Hence T possesses a unique fixed point, namely a function $F \in \mathcal{F}$ such that $TF = F$. We readily verify that the graph of F is an attractor for the IFS $(J \times K, W_i: i = 1, 2, \cdots, N)$ and so it must equal G. □

Example 1. Let $J = [0, 1]$ and let $0 = t_0 < t_1 < \cdots < t_N = 1$. Then choose $l_i(t) = t_{i-1} + (t_i - t_{i-1})t$, so that the attractor of the IFS $\{J, l_i: i = 1, 2, \cdots, N\}$ is $I = [0, 1]$. Let K be a sufficiently large bounded subset of $\mathbb{R}$ and

$$\begin{aligned}&k_i(t, X) = b_i t + a_i X + e_i, \\ &b_i = c_i(x_N - x_0) - (x_i - x_{i-1}), \\ &e_i = x_{i-1} - c_i x_0, \qquad |c_i| < 1 \quad \text{for } i = 1, 2, \cdots, N.\end{aligned}$$

Here the a_i's are the only adjustable parameters, once the interpolation points $\{(t_j, x_j): j = 0, 1, \cdots, N\}$ have been specified. The attractor of the corresponding IFS $(J \times K, W_i: i = 1, 2, \cdots, N)$, where

$$W_i\begin{pmatrix} t \\ x \end{pmatrix} = \begin{pmatrix} (t_i - t_{i-1}) & 0 \\ b_i & a_i \end{pmatrix}\begin{pmatrix} t \\ x \end{pmatrix} + \begin{pmatrix} t_{i-1} \\ e_i \end{pmatrix}$$

is the graph of a self-affine function $F: [0, 1] \to \mathbb{R}$ such that

$$F(t_j) = x_j, \qquad j = 0, 1, 2, \cdots, N.$$

An illustration is provided in Fig. 7. Such functions are considered in [B] and [BH].

Example 2. Let $J = [0, 1]$ and let $0 = t_0 < t_1 < t_2 < t_3 = 1$. Then choose $l_i(t) = t_{i-1} + (t_i - t_{i-1})t$, so that the attractor of the IFS $(J, l_i: i = 1, 2, 3)$ is $I = [0, 1]$. Let $K = [0, 1] \times [0, 1] \subset \mathbb{R}^2$ and let the maps $k_i(t, X)$ depend only on $X = (x, y)$, and each be of the form

$$k\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

where the constants a, b, c, d, e, f are fixed by requiring that

$$\begin{aligned}&k_1(0, 0) = (0, 0), \quad k_1(1, 0) = (.25, .5), \quad k_1(.5, 1) = (.5, 0), \\ &k_2(0, 0) = (.25, .5), \quad k_2(.5, 1) = (.5, 1), \quad k_2(1, 0) = (.75, .5), \\ &k_3(0, 0) = (.75, .5), \quad k_3(.5, 1) = (.5, 0), \quad k_3(1, 0) = (1, 0).\end{aligned}$$

The attractor G for the resulting IFS will be the graph of a function $F: [0, 1] \to [0, 1] \times [0, 1]$ such that $F(t_0) = (0, 0)$, $F(t_1) = (.25, .5)$, $F(t_2) = (.75, .5)$, $F(t_3) = (1, 0)$.

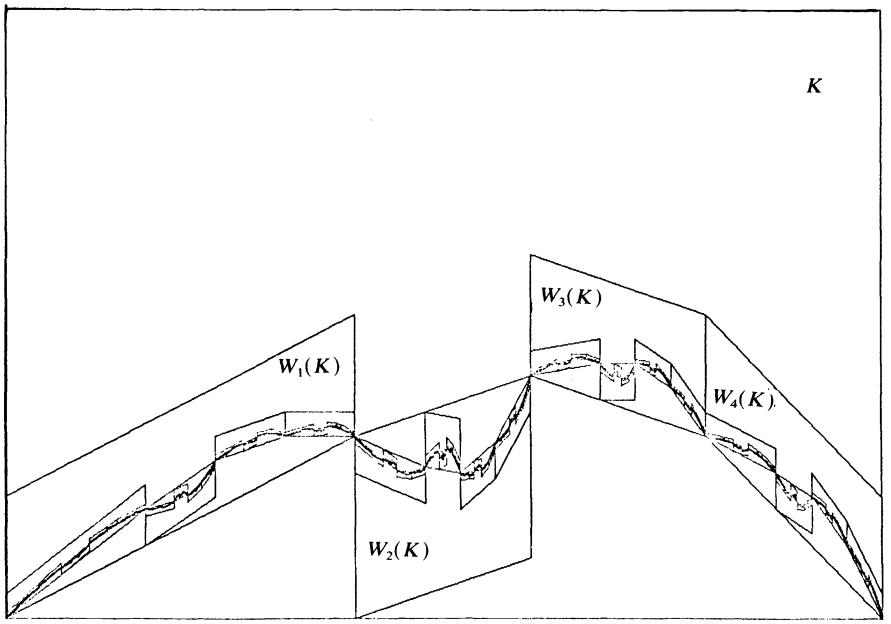


FIG. 7. *Example of a self-affine function in one dimension. The interpolation points are marked* ●, *and the vertical scaling parameters are* $a_1 = 0.2$, $a_2 = -0.3$, $a_3 = 0.2$, *and* $a_4 = 0.2$. *See Example* 1.

In particular, if we write $F(t) = (f(t), h(t))$, then $f(t)$ interpolates according to $f(t_0) = 0$, $f(t_1) = .25, f(t_2) = .75, f(t_3) = 1$. Various views of G, including the graph of f, are shown in Fig. 8. We see that G is a Sierpinski triangle when viewed from "above" the xy-plane, while it provides an intricate one-dimensional function $h(t)$ when projected onto the yt-plane. Note that this example is "diagonal" in that there is no coupling between the t and X variables. F is a map, from the elementary attractor I onto the attractor S, of a two-dimensional IFS, and is the same as the map ϕ constructed in the introduction using the code space.

Example 3. Let J here be $J \times K$ in Example 2, so that t here corresponds to (t, X) in Example 2. Let $l_i(t)$ here be $(l_i(t), k_i(t, X))$ in Example 2. Then I here is the graph $G \subset \mathbb{R}^3$ in Example 2. Let K here be $J \times K$ in Example 1, and the maps k_i here be the maps W_i in Example 1, with $N = 3$. Then the attractor for the IFS $(J \times K, W_i, i \in \{1, 2, 3\})$ obtained here will be the graph of a function that maps the entity represented in Fig. 8 onto the function shown in Fig. 7. Again, we are here dealing with a diagonal construction where the maps k_i depend only on X.

Example 4. Let $J = [0, 1]$ and let $t_0 = 0$, $t_1 = \frac{1}{2}$, and $t_2 = 1$. Choose $l_i(t) = (t_i - t_{i-1})t + t_{i-1}$. Let $K = [0, 1] \times [0, 1]$ and define maps $k_i(t, X)$: $K \to K$ by

$$k_i(t, X) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2}(-1)^{i-1} \\ \frac{1}{2}(-1)^{i-1} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} (i-1) & \frac{1}{2} \\ (i-1) & \frac{1}{2} \end{pmatrix}$$

where $X = (x, y)$ and $i \in \{1, 2\}$. Then the attractor A for the IFS $(K, k_i, i \in \{1, 2\})$ is a Peano Curve. Figure 9 shows A and also different views of G, the attractor of the IFS $(J \times K, W_i, i \in \{1, 2, 3\})$ with $W_i(t, X) = (l_i(t), k_i(t, X))$. Note that k_i is again diagonal in that it is not coupled to t.

Example 5. Let $J = [0, 1]$ and $K = [0, 1] \times [0, 1]$. Let $\{X_j : j \in \{0, \cdots, N\}\}$ and $\{(t_j, \Theta_j) : j \in \{0, \cdots, N\}\}$ be given sets of distinct data points in K and $J \times K$, respectively, with $0 = t_0 < \cdots < t_N = 1$.

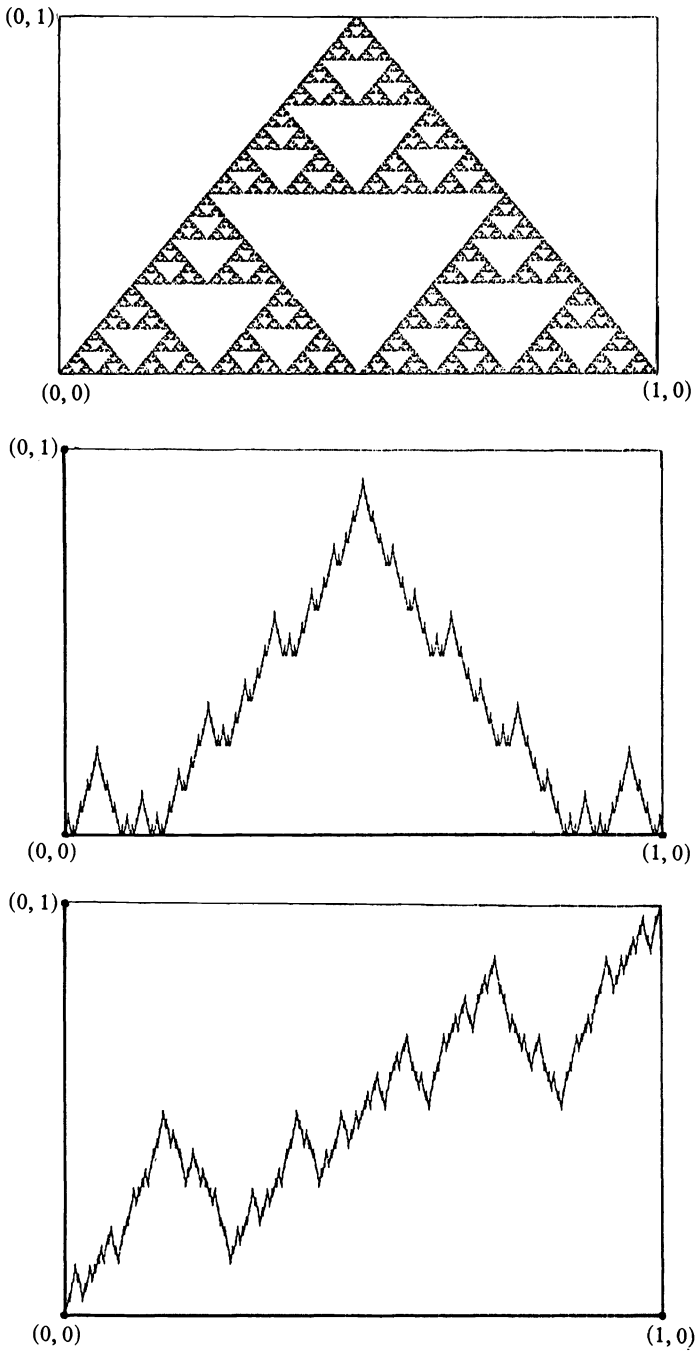


FIG. 8. *Several views of the attractor G for the* IFS *in Example* 2.

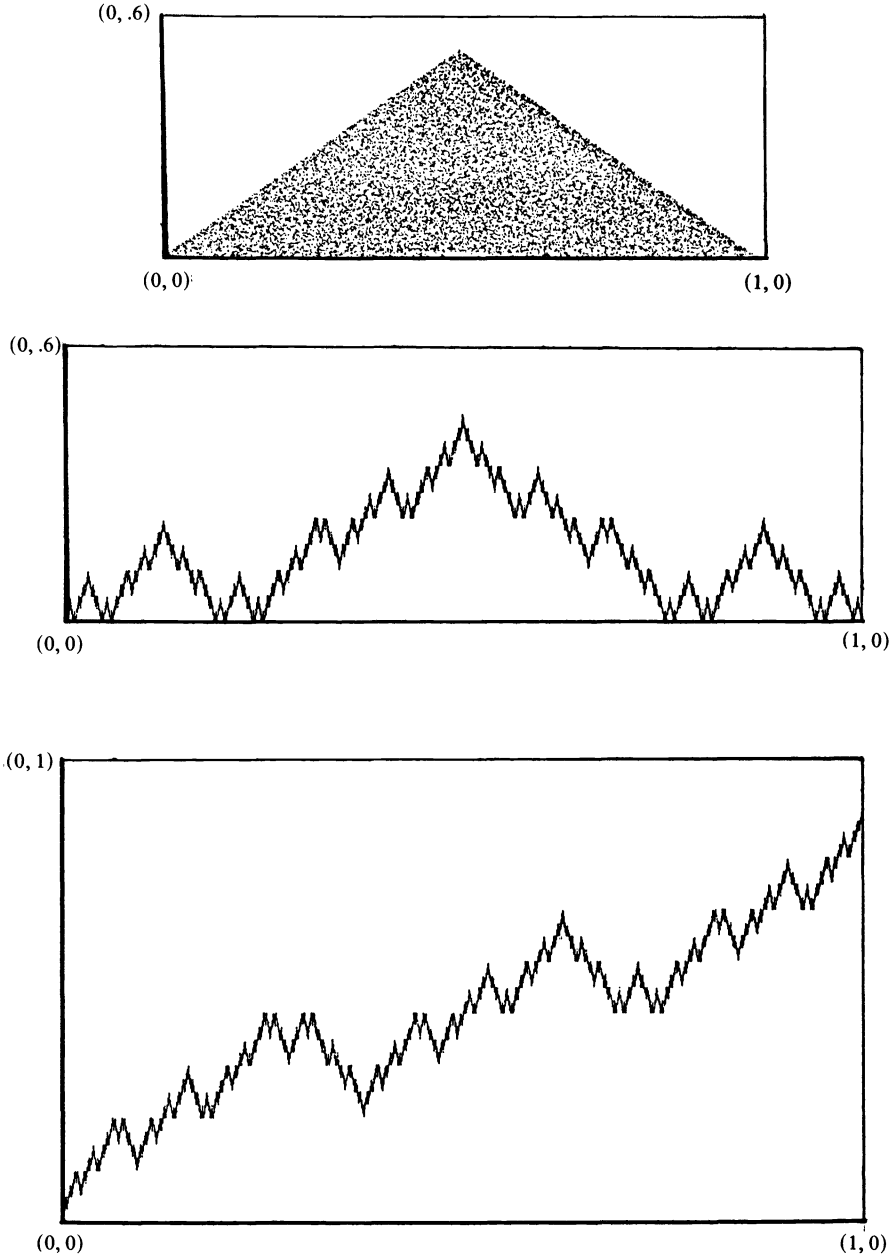


FIG. 9. *Several views of the attractor G for the* IFS *in Example* 4.

Define maps $l_i : J$ by

$$l_i(t) = (t_i - t_{i-1})t + t_{i-1}$$

and k_i: $J \times K \to K$ by

$$k_i(t, X) = \begin{pmatrix} \Delta\xi_i - \Delta x_i & \Delta x_i & -\Delta y_i \\ \Delta\eta_i - \Delta y_i & \Delta y_i & \Delta x_i \end{pmatrix} \begin{pmatrix} t \\ x \\ y \end{pmatrix} + \begin{pmatrix} \xi_{i-1} \\ \eta_{i-1} \end{pmatrix}$$

with $X_j = (x_j, y_j)$ and $\Theta_j = (\xi_j, \eta_j)$, for all $j \in \{0, \cdots, N\}$.

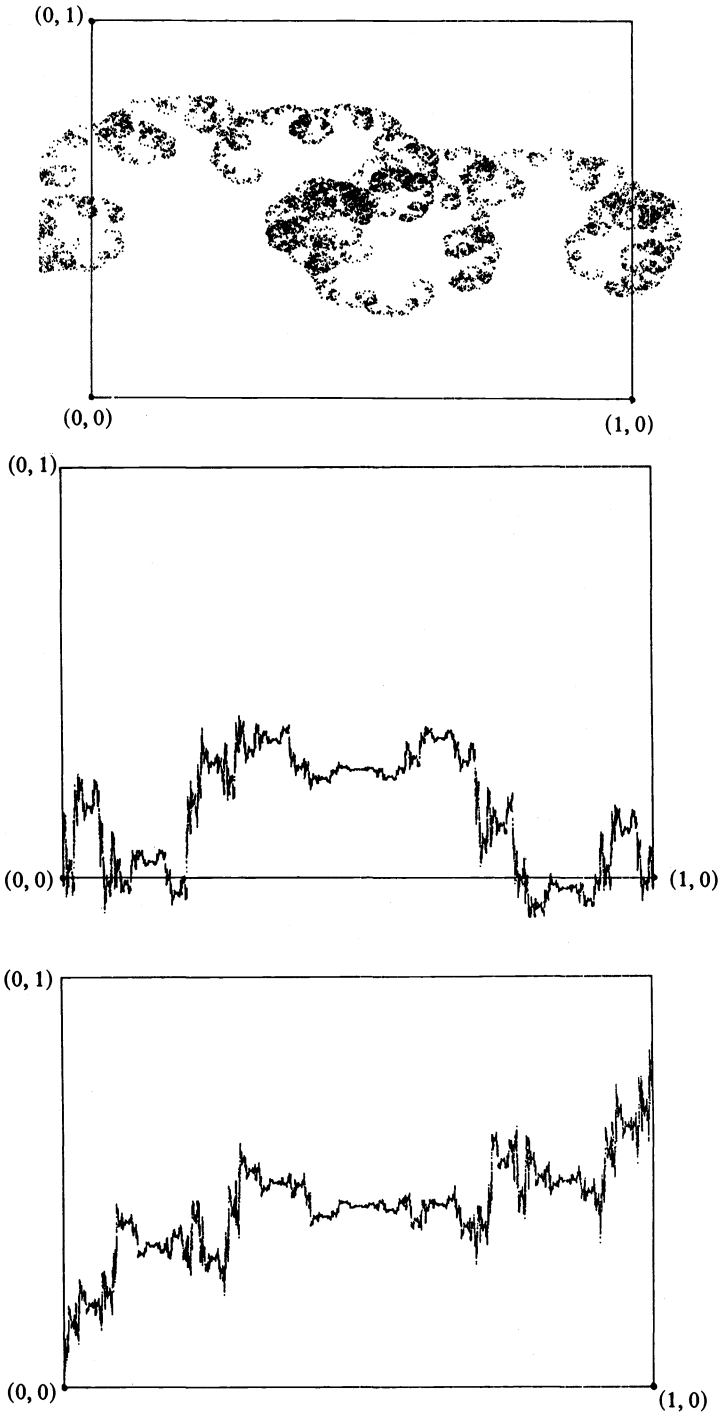


FIG. 10. *Views of the projections of the attractor for the* IFS *in Example* 5.

Then the attractor G of the IFS $(J \times K, W_i, i \in \{1, \cdots, N\})$, where $W_i(t, X) = (l_i(t), k_i(t, X))$, is the graph of a continuous function containing $\{(t_j, \Theta_j): j \in \{0, \cdots, N\}\}$.

Note that if $\Theta_j = X_j$, for all $j \in \{0, \cdots, N\}$, then the k_i are maps as considered in Example 2 and the projection of G onto K is the attractor A for the IFS

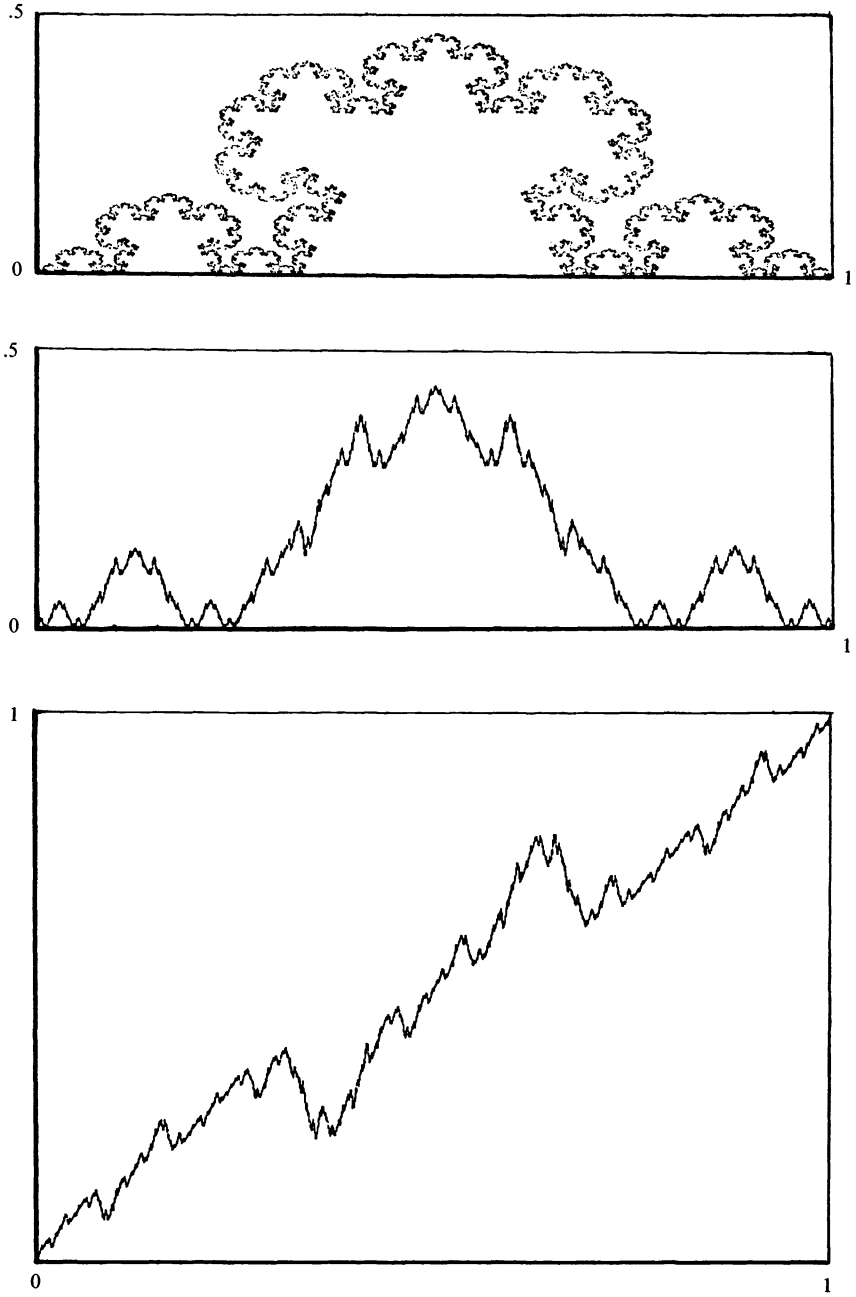


FIG. 11. *Several views of the attractor for the* IFS *in Example* 6.

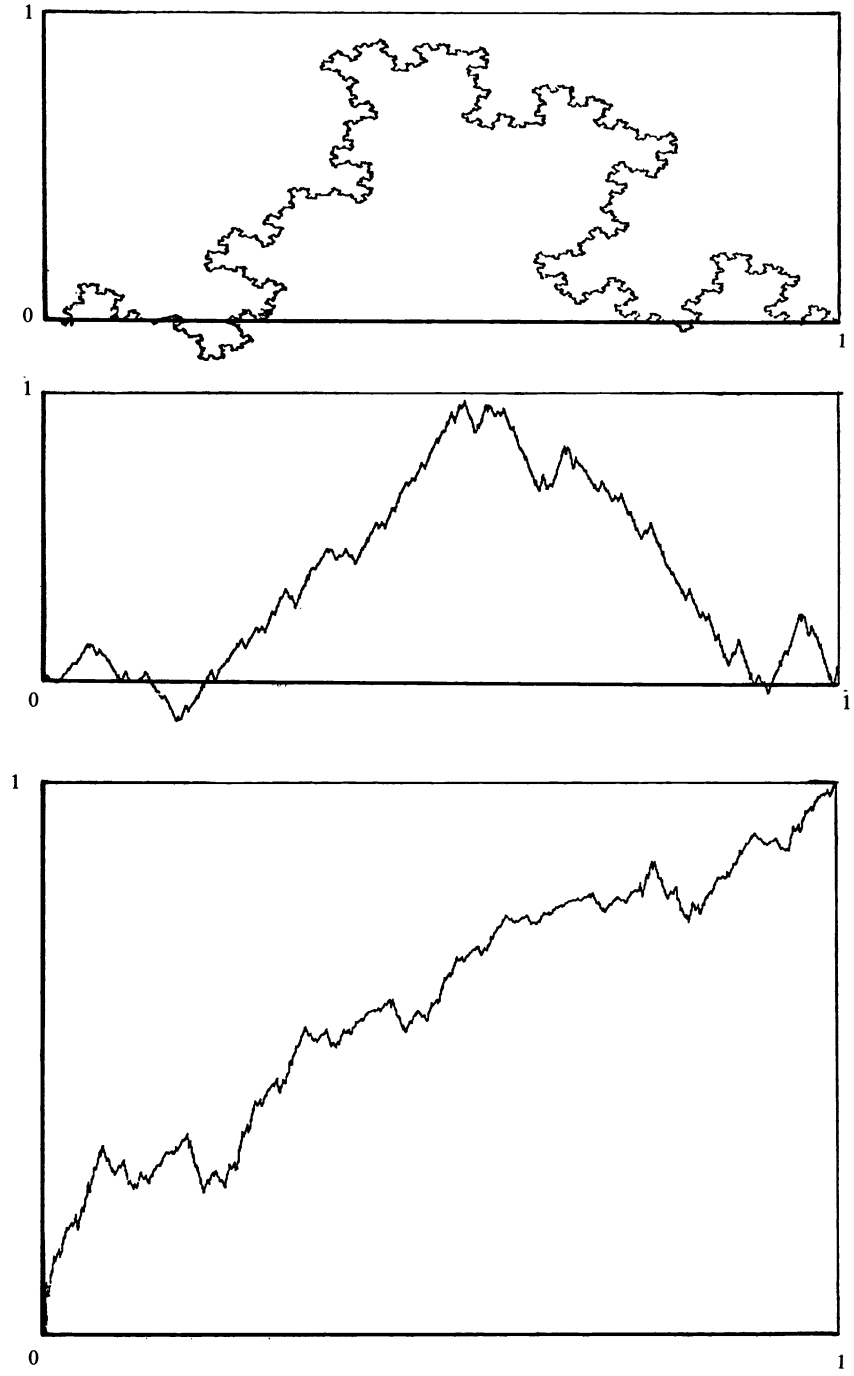


FIG. 12. *Views of the attractor G for an* IFS *of the form as in Example* 7.

$(K, k_i, i \in \{1, \cdots, N\})$, where $k_i : K \to K$:

$$k_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Delta x_i & -\Delta y_i \\ \Delta y_i & \Delta x_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix}.$$

Figure 10 shows different views of G for $N = 4$, $\{X_j : j \in \{0, \cdots, 4\}\} = \{(0, 0), (\frac{1}{4}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{4}), (\frac{3}{4}, \frac{1}{2}), (1.0)\}$ and $\{(t_j, \Theta_j) : j \in \{0, \cdots, 4\}\} = \{(0, 0, 0), (\frac{1}{4}, \frac{3}{10}, \frac{3}{5}), (\frac{55}{100}, \frac{3}{5}, \frac{3}{10}), (\frac{7}{10}, \frac{1}{2}, \frac{2}{5}), (1, 1, \frac{1}{5})\}$.

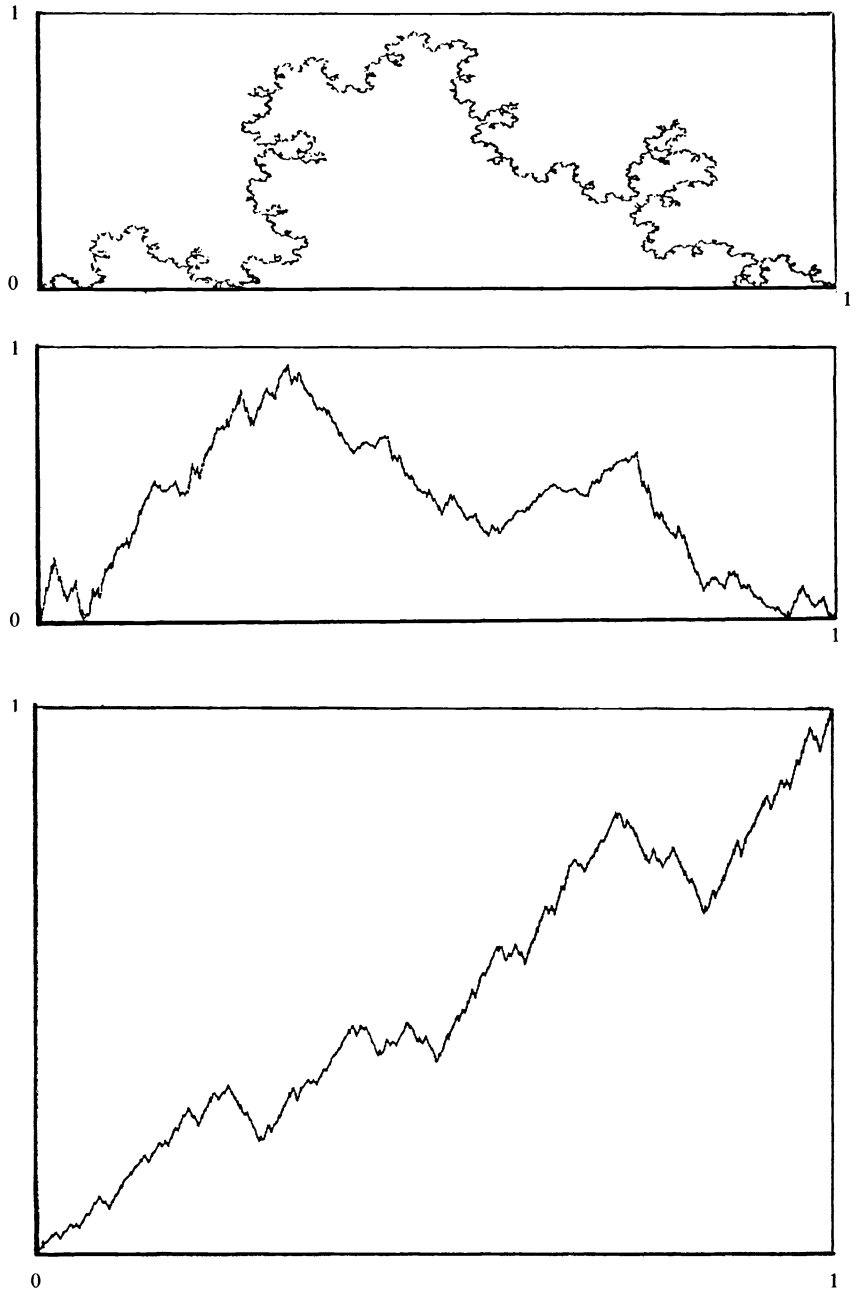


FIG. 13. *Views of the attractor for an* IFS *of the form as in Example* 7.

Example 6. Let $J=[0,1]$ and $K=[0,1]\times[0,1]$. Choose maps $W_i(t,X): J\times K$ with

$$W_i(t,X)=\begin{pmatrix}1/N & 0 & 0\\ 0 & s_i\cos\theta_i & -s_i\sin\theta_i\\ 0 & s_i\sin\theta_i & s_i\cos\theta_i\end{pmatrix}(t,X)+\begin{pmatrix}(i-1)/N\\ \alpha_i\\ \beta_i\end{pmatrix},$$

$i\in\{1,\cdots,N\}$. Let $N=4$. Determine s_i, θ_i, α_i, and β_i such that $W_1(0,0,0)=(0,0,0)$, $W_4=(1,1,0)=(1,1,0)$, $W_2(0,0,0)=W_1(1,1,0)=(\frac{1}{4},\frac{1}{3},0)$, $W_3(0,0,0)=W_2(1,1,0)=(\frac{1}{2},\frac{1}{2},\sqrt{3}/4)$, $W_4(0,0,0)=W_3(1,1,0)=(\frac{3}{4},\frac{2}{3},0)$. Figure 11 shows views of the attractor G of the IFS $(J\times K, W_i, i\in\{1,\cdots,4\})$.

Example 7. Let J, K be as in Example 6. Let maps $W_i: J\times K$ be defined as in Example 6 above. Determine the parameters s_i, θ_i, α_i, and β_i according to (2.7) for some set of interpolation points $\{(t_j,X_j): j\in\{1,\cdots,4\}\}$. Figure 12 shows three views of the corresponding attractor, for an arbitrarily chosen set of parameter values. Figure 13 corresponds to a different set of parameter values.

3. Fractal dimension of graphs of self-affine functions.

3.1. Definition of fractal dimension and related topics. Let E be a bounded set in $\mathbb{R}^n$ and for $\varepsilon>0$ let $\mathcal{N}(\varepsilon)$ be the minimum number of balls of diameter ε necessary to cover E. If

$$\lim_{\varepsilon\to 0}\frac{\log\mathcal{N}(\varepsilon)}{\log 1/\varepsilon}\equiv\dim(E) \tag{3.1}$$

exists, then it is called the (fractal) dimension (also known as the capacity or box-counting dimension) of E. Note that if E is a surface of finite area A then $\mathcal{N}(\varepsilon)\propto A\varepsilon^{-2}$ and thus $\dim(E)=2$ or if E is a curve of finite length L then $\mathcal{N}(\varepsilon)\propto L\varepsilon^{-1}$ and hence $\dim(E)=1$.

Sometimes it is useful to consider covers by sets other than ε-balls. Let $\{\mathcal{C}_\varepsilon: \varepsilon>0\}$ be a family of covers of E and let $\mathcal{N}^*(\varepsilon)$ be the minimum number of sets $C\in\mathcal{C}_\varepsilon$ needed to cover E. Suppose that there exist positive constants c_1 and c_2 so that

$$c_1\mathcal{N}(\varepsilon)\leqq\mathcal{N}^*(\varepsilon)\leqq c_2\mathcal{N}(\varepsilon)$$

then it is clear that $\mathcal{N}(\varepsilon)$ in (3.1) can be replaced by $\mathcal{N}^*(\varepsilon)$. Another notion of dimension is Hausdorff dimension $\mathrm{HD}(E)$:

$$\mathrm{HD}(E)=\sup\{0\leqq d<\infty: H_d(E)>0\}$$

with

$$H_d(E)=\lim_{\varepsilon\to 0}\left(\inf_{\mathcal{U}}\sum_i|U_i|^d\right)$$

where the inf is taken over all countable covers $\mathcal{U}$ of E such that $|U_i|\leqq\varepsilon$ and $|U_i|=\sup\{|x-y|: x,y\in U_i\}$.

Note that

$$\mathrm{HD}(E)\leqq\liminf_{\varepsilon\to 0}\frac{\log\mathcal{N}(\varepsilon)}{\log 1/\varepsilon}.$$

Let $(K, w_i, i=1,\cdots,N)$ be an IFS with attractor A and suppose that:

(3.2) (a) $w_i: K\to K$ is a similitude, i.e., $d(w_i(x),w_i(y))=s_id(x,y)$, for some $0\leqq s_i<1$ for all i;

(b) There exists a nonempty open set O such that $\cup w_i(O) \subseteq O$ and $w_i(O) \cap w_j(O) = \phi$ for $i \neq j$.

Then a result of [Mo] (similar results are found in [H] and [BD]) shows that HD (A) is the unique positive solution of $\sum_{i=1}^{N} s_i^d = 1$. We now show that under conditions (3.2), dim (A) = HD (A).

PROPOSITION 1. *Let* $\{K, w_i, i = 1, \cdots, N\}$ *be an* IFS *satisfying* (3.2). *If* A *is the attractor of this* IFS, *then* HD (A) = dim (A).

Proof. For $\varepsilon > 0$ let $\mathcal{N}(\varepsilon)$ ($\mathcal{N}_i(\varepsilon)$, $i = 1, \cdots, N$) be the minimum number of ε-balls required to cover A ($w_i(A)$, $i = 1, \cdots, N$), respectively. Since w_i is a similitude we have $\mathcal{N}_i(\varepsilon) = \mathcal{N}(\varepsilon / s_i)$ and thus

$$\mathcal{N}(\varepsilon) \leqq \sum_{i=1}^{N} \mathcal{N}_i(\varepsilon) = \sum_{i=1}^{N} \mathcal{N}\left(\frac{\varepsilon}{s_i}\right).$$

Let $\underline{s} = \min\{s_i: i = 1, \cdots, N\}$ and $\bar{s} = \max\{s_i: i = 1, \cdots, N\}$. Let D be the unique positive solution of $\sum_{i=1}^{N} s_i^D = 1$ and choose $c > 0$ such that

$$\mathcal{N}(\varepsilon) \leqq c\varepsilon^{-D} \quad \text{for } \underline{s} \leqq \varepsilon \leqq 1. \tag{3.3}$$

Suppose $\mathcal{N}(\varepsilon) \leqq c\varepsilon^{-D}$ for $(\bar{s})^n \underline{s} \leqq \varepsilon \leqq 1$. Then if $(\bar{s})^{n+1} \underline{s} \leqq \varepsilon \leqq \underline{s}$ we have $(\bar{s})^n \underline{s} \leqq \varepsilon / s_i \leqq 1$ and so $\mathcal{N}(\varepsilon) \leqq \sum_{i=1}^{N} \mathcal{N}(\varepsilon / s_i) \leqq \sum_{i=1}^{N} c s_i^D \varepsilon^{-D} = c\varepsilon^{-D}$. By induction

$$\mathcal{N}(\varepsilon) \leqq c\varepsilon^{-D} \quad \text{for } 0 < \varepsilon \leqq 1$$

and so

$$\limsup_{\varepsilon \to 0} \frac{\log \mathcal{N}(\varepsilon)}{\log 1/\varepsilon} \leqq D.$$

But also, as previously noted,

$$\liminf_{\varepsilon \to 0} \frac{\log \mathcal{N}(\varepsilon)}{\log 1/\varepsilon} \geqq D$$

and so dim $(A) = D$. □

3.2. Dimension of self-affine functions. We now consider dim (G) (recall the notation of § 2). We restrict our attention to the following situation:

(a) $k_i: K \to K$, i.e., no coupling between J and K;

(b) $d(l_i(\tau_1), l_i(\tau_2)) \leqq s_i d(\tau_1, \tau_2) \quad \forall \tau_1, \tau_2 \in J,$

$d(k_i(\xi_i), k_i(\xi_2)) \leqq s_i d(\xi_1, \xi_2) \quad \forall \xi_1, \xi_2 \in K \quad$ where $0 < s_i < 1 \quad \forall i = 1, \cdots, N.$

THEOREM 2. *Let* $F: I \to K$ *be as in* § 2, *and let* D *be the unique positive solution of* $\sum_{i=1}^{N} s_i^D = 1$. *Then* dim $(G) \leqq D$.

Proof. Let $\mathcal{N}(\varepsilon)$ ($\mathcal{N}_i(\varepsilon)$) be the minimum number of ε-balls needed to cover G ($W_i(G)$). Since W_i maps any (ε / s_i)-ball into an ε-ball we have $\mathcal{N}_i(\varepsilon) \leqq \mathcal{N}(\varepsilon / s_i)$ and thus

$$\mathcal{N}(\varepsilon) \leqq \sum_{i=1}^{N} \mathcal{N}_i(\varepsilon) \leqq \sum_{i=1}^{N} \mathcal{N}\left(\frac{\varepsilon}{s_i}\right).$$

As in Proposition 1 this implies dim $(G) \leqq D$. □

We now consider the special case where the IFS $\{K, k_i, i = 1, \cdots, N\}$ satisfies the conditions of (3.2). In particular, we have Example 6 in mind; however, we will remain slightly more general here. Let D be the unique positive solution of $\sum_{i=1}^{N} s_i^D = 1$. By

Theorem 2, $\dim(G) \leqq D$. However, as $A = \operatorname{proj}_K G$ we have $\dim(G) \geqq \dim(A) = D$ and thus we have Corollary 1.

COROLLARY 1. $\dim(G) = D$.

Next consider the "equal scaling" case of Example 2.6: $l_i(t) = t/N + (i-1)/N$, for all $i = 1, \cdots, N$. We can relax the conditions $s_i \geqq 1/N$ in Corollary 1.

THEOREM 3. $\dim(G) = D$, *where* D *is the unique positive solution of* $\sum_{i=1}^{N} s_i^D = 1$.

Proof. For simplicity we assume all s_i's are nonzero. Let $\mathscr{C}(n) = \{[(i-1)/N^n, i/N^n] \times B: i = 1, \cdots, N^n$, with B a $1/n$-ball in $K\}$. Let $\mathscr{N}^*(n)$ be the minimum number of sets from $\mathscr{C}(n)$ and let $\mathscr{N}(\varepsilon)$ and $\mathscr{N}_p(\varepsilon)$ be the minimum number of ε-balls necessary to cover G and $A = \operatorname{proj}_K G$, respectively. Note that $\mathscr{N}^*(n) \geqq \mathscr{N}(2/N^n)$. Let

$$G_i = G|_{[(i-1)/N^n, i/N^n] \times K}$$

and let $\omega_1 \omega_2 \cdots \omega_n$, where $\omega_i \in \{1, \cdots, N\}$, be the code corresponding to $[(i-1)/N^n, i/N^n]$. Let $\mathscr{B}$ be a minimal $(1/N^n \prod_{i=1}^{n} s_{\omega_i})$-cover of A. By applying $W_{\omega_1} \cdots W_{\omega_n}$ to $\{[0,1] \times B \mid B \in \mathscr{B}\}$ we can cover G_i with

$$\mathscr{N}_p\left(\frac{1}{N^n \prod_{i=1}^{n} s_{\omega_i}}\right)$$

sets from $\mathscr{C}(n)$. Thus

$$\begin{aligned}
\mathscr{N}^*(n) &\leqq \sum_{\omega_1=1}^{N} \cdots \sum_{\omega_n=1}^{N} \mathscr{N}_p\left(\frac{1}{N^n \prod_{i=1}^{N} s_{\omega_i}}\right) \\
&\leqq N^{nD} c \sum_{\omega_1=1}^{N} \cdots \sum_{\omega_n=1}^{N} \prod_{i=1}^{N} s_{\omega_i}^D \\
&= N^{nD} c \left(\sum_{i=1}^{N} s_i^D\right) = c N^{nD}
\end{aligned}$$

by (3.3). Thus for $2/N^n \leqq \varepsilon \leqq 2/N^{n-1}$ we have

$$\frac{\log \mathscr{N}(\varepsilon)}{\log 1/\varepsilon} \leqq \frac{\log(cN^{nD})}{\log(N^{n-1}/2)}$$

and so

$$\limsup_{\varepsilon \to 0} \frac{\log \mathscr{N}(\varepsilon)}{\log 1/\varepsilon} \leqq D.$$

As $\dim(A) = D$ we have the desired result. □

4. Fractal dimension of projections of self-affine functions.

4.1. The dimension of one-dimensional self-affine functions. In this section we consider the dimension of the graphs of the functions constructed in Example 1. (We use the notation of Example 1 here.) We show that if

$$\sum_{i=1}^{N} |a_i| > 1 \quad \text{and} \quad \{(t_j, x_j): j = 0, \cdots, N\} \text{ is not collinear,} \tag{4.1}$$

then $\dim(G)$ is given by

$$\sum_{i=1}^{N} |a_i| b_i^{\dim(G)-1} = 1. \tag{4.2}$$

Let $h(d)=\sum_{i=1}^{N}|a_i|b_i^{d-1}$. Since $h(d)$ is strictly decreasing, $\lim_{d\to-\infty}h(d)=\infty$, and $h(2)\leqq\max_{i=1,\cdots,N}|a_i|<1$, there is a unique $D\in(-\infty,2)$ such that $h(D)=1$. Also $D>1$ if and only if $\sum_{i=1}^{N}|a_i|>1$; thus (4.2) uniquely determines $\dim(G)\in(1,2)$.

Let us now define a class of covers that allows us to relate covers of different sizes.

DEFINITION. For $0<\varepsilon<1$, $\{\tau_l\}_{l=0}^{m}$ is called an *ε-partition* if

(a) $\tau_l\in(-\varepsilon/2,1)$,

(b) $\varepsilon/2<\tau_{l+1}-\tau_l\leqq\varepsilon$,

for $l=1,2,\cdots,n-1$. A cover $\mathscr{C}$ of G will be called an *ε-column cover of G* with associated ε-partition $\{\tau_l\}_{l=0}^{m}$ if there are positive integers $n_0,\cdots,n_m$ and real numbers $\xi_0,\cdots,\xi_m$ such that

$$\mathscr{C}=\{[\tau_k,\tau_k+\varepsilon]\times[\xi_k+(j_k-1)\varepsilon,\xi_k+j_k\varepsilon]:j_k=1,\cdots,n_k;\ k=0,1,\cdots,m\}.$$

Note that $\mathscr{C}$ consists of $\sum_{k=0}^{m}n_k$ closed $\varepsilon\times\varepsilon$ squares arranged in $m+1$ columns. Let $|\mathscr{C}|$ denote the cardinality of $\mathscr{C}$ and define $\mathscr{N}^*(\varepsilon)=\min\{|\mathscr{C}|:\mathscr{C}$ is an ε-column cover of $G\}$ and let $\mathscr{N}(\varepsilon)$ be the minimum number of $\varepsilon\times\varepsilon$ squares $[a,a+\varepsilon]\times[b,b+\varepsilon]$, $a,b\in\mathbb{R}$, required to cover G. Lemma 4.1 below shows that $\mathscr{N}^*(\varepsilon)$ can be used in the calculation of $\dim(G)$.

LEMMA 4.1. *$\mathscr{N}(\varepsilon)\leqq\mathscr{N}^*(\varepsilon)\leqq2\mathscr{N}(\varepsilon)$, for all $0<\varepsilon<1$.*

Proof. Clearly, $\mathscr{N}(\varepsilon)\leqq\mathscr{N}^*(\varepsilon)$.

We introduce a third class of covers: a cover $\mathscr{C}$ of G will be called an *ε-nonoverlapping cover of G* if it consists of $\varepsilon\times\varepsilon$ squares with nonintersecting interiors of the form $[k\varepsilon,(k+1)\varepsilon]\times[y,y+\varepsilon]$, where $k\in\{0,1,\cdots,[\![1/\varepsilon]\!]\}$ and $y\in\mathbb{R}$. Define $\mathscr{N}^{**}(\varepsilon)=\min\{|\mathscr{C}|:\mathscr{C}$ is an ε-nonoverlapping cover of $G\}$.

Clearly, $\mathscr{N}^{**}(\varepsilon)\leqq2\mathscr{N}(\varepsilon)$. If $\mathscr{C}$ is a minimal ε-nonoverlapping cover of G then, since G is the graph of a continuous function, $\mathscr{C}$ is also an ε-column cover of G. Thus $\mathscr{N}^*(\varepsilon)\leqq\mathscr{N}^{**}(\varepsilon)\leqq2\mathscr{N}(\varepsilon)$. □

In the proof of Theorem 4 we show that $\mathscr{N}(\varepsilon)$ satisfies the functional inequality

$$(*)\qquad \sum_{i=1}^{N}\frac{|a_i|}{b_i}\mathscr{N}^*\left(\frac{\varepsilon}{b_i}\right)-\frac{\beta_1}{\varepsilon}\leqq\mathscr{N}^*(\varepsilon)\leqq\sum_{i=1}^{N}\frac{|a_i|}{b_i}\mathscr{N}^*\left(\frac{\varepsilon}{b_i}\right)+\frac{\beta_2}{\varepsilon}$$

for some $\beta_1,\beta_2>0$ and all $0<\varepsilon<1$.

To show that the $1/\varepsilon$ term becomes negligible we need the following lemma.

LEMMA 4.2. *If (4.1) is satisfied, then*

$$\lim_{\varepsilon\to0}\varepsilon\mathscr{N}^*(\varepsilon)=\infty.$$

Proof. As $\{(t_j,x_j):j=0,\cdots,N\}$ is not collinear there is some $l\in[1,\cdots,N-1]$ such that $V\equiv|t_l-(t_N-t_0)x_l-t_0|>0$. Let $\underline{b}=\min\{b_i:i=1,\cdots,N\}$. Since F is continuous we obtain

$$\begin{aligned}N(\varepsilon)&\geqq\sum_{i_1=1}^{N}\cdots\sum_{i_k=1}^{N}\left(|a_{i_1}\cdots a_{i_k}|\frac{V}{\varepsilon}\right)\\&\geqq\left(\sum_{i=1}^{N}|a_i|\right)^k\frac{V}{\varepsilon}\quad\text{for }0<\varepsilon<\underline{b}^k\text{ and }k\in\mathbb{N}.\end{aligned}$$

Since $(\sum_{i=1}^{N}|a_i|)>1$ the lemma is proved. (See [HM] for more detail.) □

The investigation of the following theorem commenced in [BH] and [HM].

THEOREM 4. *Let F be the function with graph G generated by the* IFS *$(K,W_i,i=1,\cdots,N)$, where*

$$W_i\begin{pmatrix}t\\x\end{pmatrix}=\begin{pmatrix}b_i&0\\c_i&a_i\end{pmatrix}\begin{pmatrix}t\\x\end{pmatrix}+\begin{pmatrix}t_{i-1}\\x_{i-1}\end{pmatrix}\quad\forall i=1,\cdots,N$$

with a_i, b_i, c_i, t_i, x_i *as in Example* 1. *If* $\sum_{i=1}^{N}|a_i|>1$ *and* $\{(t_j, x_j): j=0, 1, \cdots, N\}$ *is not collinear, then* $\dim(G)$ *is the unique real solution* D *of* $\sum_{i=1}^{N}|a_i|b_i^{D-1}=1$; *otherwise* $\dim(G)=1$.

Proof. We first obtain functional inequalities for $\mathcal{N}^*(\varepsilon)$. Let $0<\varepsilon<1$ and let $\mathscr{C}$ be a minimal ε-column cover of G with associated ε-partition $\{\tau_j\}_{j=0}^{m}$. For $i\in\{1, \cdots, N\}$ let $[a, b+\varepsilon]$, $a, b\in\mathbb{R}$, be the smallest interval of the form $[\tau_k, \tau_l+\varepsilon]$ that covers $[t_{i-1}, t_i]$. Let

$$\mathscr{C}_i=\{C\in\mathscr{C}: C\subset[\tau_k, \tau_l+\varepsilon]\times\mathbb{R}\}$$

and $\mathcal{N}_i=|\mathscr{C}_i|$. Since F is continuous, $\max_{t\in[0,1]}|F(t)|=M<\infty$. Since there are at most two columns in $\mathscr{C}_i\cap\mathscr{C}_{i+1}$ there is some $\alpha_1>0$ such that $\sum_{i=1}^{N}\mathcal{N}_i\leqq\mathcal{N}^*(\varepsilon)+\alpha_1/\varepsilon$ for $0<\varepsilon<1$.

Suppose $a_i\neq 0$. Then W_i is invertible. Consider a typical column R in $\mathscr{C}_i$ that consists of n $\varepsilon\times\varepsilon$ squares; then $W_i^{-1}(R)$ is a parallelogram that can be covered by

$$\left[\!\left[\frac{nb_i}{|a_i|}+\left|\frac{c_i}{b_i}\right|+1\right]\!\right]$$

squares of side ε/b_i as in Fig. 14. Applying W_i^{-1} generates an (ε/b_i)-column cover $\mathscr{D}_i$ of G. Since there are at most $2b_i/\varepsilon+2$ columns in $\mathscr{C}_i$ there is some $\beta_1>0$ so that

$$\mathcal{N}^*(\varepsilon)\geqq\sum_{i=1}^{N}\mathcal{N}_i-\frac{b_1}{\varepsilon}\geqq\sum_{i=1}^{N}\frac{|a_i|}{b_i}\mathcal{N}^*\left(\frac{\varepsilon}{b_i}\right)-\frac{c_1}{\varepsilon}.$$

Next we obtain a lower bound for $\mathcal{N}^*(\varepsilon)$. Let $\mathscr{D}_i$ be a best ε/b_i-column cover of G and R a typical column of $\mathscr{D}_i$. Note that $W_i(R)$ is a parallelogram that can be covered by

$$\left[\!\left[n\frac{|a_i|}{b_i}+\frac{|c_i|}{b_i}+1\right]\!\right]$$

$\varepsilon\times\varepsilon$ squares. In this way we generate a cover $\mathscr{C}_i$ of $W_i(G)$ consisting of $\varepsilon\times\varepsilon$ squares. As there are at most $[(2b_i/\varepsilon)+2]$ columns of $\mathscr{D}_i$ there is some $\alpha_2>0$ such that

$$|\mathscr{C}_i|\leqq\frac{|a_i|}{b_i}\mathcal{N}^*\left(\frac{\varepsilon}{b_i}\right)+\frac{\alpha_2}{\varepsilon}\quad\text{for } 0<\varepsilon<1.$$

Note that $\cup\,\mathscr{C}_i$ may not be an ε-column cover of G because the columns of $\mathscr{C}_i$ may not join up properly with those of $\mathscr{C}_{i+1}$; however, an ε-column cover $\mathscr{C}$ can be constructed from $\cup\,\mathscr{C}_i$ by replacing at most two columns from $\mathscr{C}_i\cup\mathscr{C}_{i+1}$ with at most two properly spaced columns, and so

$$\mathcal{N}^*(\varepsilon)\leqq|\mathscr{C}|\leqq\sum_{i=1}^{N}|\mathscr{C}_i|+(N)\left(\frac{4n}{\varepsilon}+1\right).$$

Thus there is $\beta_2>0$ such that

$$\mathcal{N}^*(\varepsilon)\leqq\sum_{i=1}^{N}\frac{|a_i|}{b_i}\mathcal{N}^*\left(\frac{\varepsilon}{b_i}\right)+\frac{\beta_2}{\varepsilon}$$

and so we have established (*).

Case 1. Condition (4.1) is satisfied.

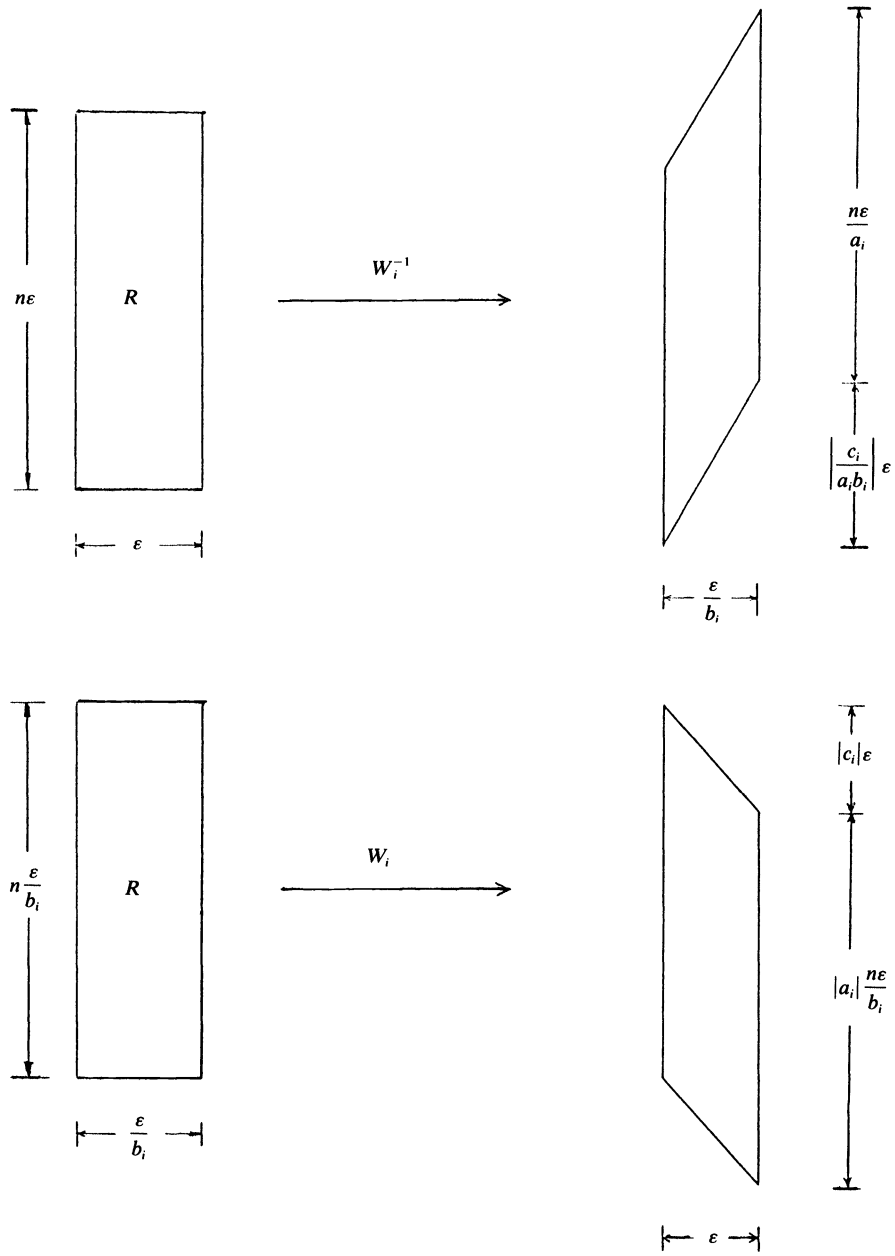


FIG. 14. *The rectangle $\mathcal{R}$ and its images under the maps W_i and W_i^{-1}. See the proof of Theorem* 4.

Here we will show that there exist positive constants B_1, B_2, and ε_0 so that $B_1\varepsilon^{-D} < \mathcal{N}^*(\varepsilon) < B_2\varepsilon^{-D}$ for $0 < \varepsilon < \varepsilon_0$. (Recall that D is the unique solution of $h(D) = 1$ and that $D \in (1, 2)$.)

Let $\underline{b} = \min\{b_i: i = 1, \cdots, N\}$, $\bar{b} = \max\{b_i: i = 1, \cdots, N\}$, and $\gamma = \sum_{i=1}^{N} |a_i| > 1$. By Lemma 2.1 we can select $\varepsilon_0 > 0$ small enough so that

$$\mathcal{N}^*(\varepsilon) \geqq 2\left(\frac{\beta_1}{\gamma - 1}\right)\varepsilon^{-1} \quad \text{for } 0 < \varepsilon \leqq \frac{\varepsilon_0}{\underline{b}}.$$

Pick $k_1 > 0$ small enough so that $k_1 \varepsilon_0^{-D} \leqq (\beta_1/(\gamma-1))\varepsilon_0^{-1}$ and select k_2 large enough so that

$$\mathcal{N}^*(\varepsilon) \leqq \frac{\beta_2}{1-\gamma}\varepsilon^{-1} + k_2\varepsilon^{-D} \quad \text{for } \varepsilon_0 \leqq \varepsilon \leqq \frac{\varepsilon_0}{\underline{b}}.$$

Define

$$\underline{\phi}(\varepsilon) = \left(\frac{\beta_1}{\gamma-1}\right)\varepsilon^{-1} + k_1\varepsilon^{-D},$$

$$\bar{\phi}(\varepsilon) = \left(\frac{\beta_2}{1-\gamma}\right)\varepsilon^{-1} + k_2\varepsilon^{-D}.$$

Then we have

$$\underline{\phi}(\varepsilon) \leqq \mathcal{N}^*(\varepsilon) \leqq \bar{\phi}(\varepsilon) \tag{4.3}$$

for $\varepsilon_0 \leqq \varepsilon \leqq \varepsilon_0/\underline{b}$. Note that

$$\underline{\phi}(\varepsilon) = \sum_{i=1}^{N} \frac{|a_i|}{b_i}\phi\left(\frac{\varepsilon}{b_i}\right) - \frac{\beta_1}{\varepsilon},$$

$$\bar{\phi}(\varepsilon) = \sum_{i=1}^{N} \frac{|a_i|}{b_i}\bar{\phi}\left(\frac{\varepsilon}{b_i}\right) + \frac{\beta_2}{\varepsilon}.$$

If $\bar{b}\varepsilon_0 \leqq \varepsilon \leqq \varepsilon_0$ then $\varepsilon_0 \leqq \varepsilon/b_i \leqq \varepsilon_0/\underline{b}$ so that

$$\begin{aligned}\mathcal{N}^*(\varepsilon) &\leqq \sum_{i=1}^{N} \frac{|a_i|}{b_i}\mathcal{N}^*\left(\frac{\varepsilon}{b_i}\right) + \frac{\beta_2}{\varepsilon} \\ &\leqq \sum_{i=1}^{N} \frac{|a_i|}{b_i}\bar{\phi}\left(\frac{\varepsilon}{b_i}\right) + \frac{\beta_2}{\varepsilon} = \bar{\phi}(\varepsilon)\end{aligned}$$

and in the same way $\mathcal{N}^*(\varepsilon) \geqq \underline{\phi}(\varepsilon)$ for $\bar{b}\varepsilon_0 \leqq \varepsilon \leqq \varepsilon_0$.

Suppose (4.3) holds for $\bar{b}^n\varepsilon_0 \leqq \varepsilon \leqq \varepsilon_0$; then, as in the preceding argument, (4.3) must hold for $\bar{b}^{n+1}\varepsilon_0 \leqq \varepsilon \leqq \varepsilon_0$. By induction (4.3) holds for $0 < \varepsilon \leqq \varepsilon_0$. Since $D > 1$ there exist positive constants B_1 and B_2 such that $B_2\varepsilon^{-D} \leqq \underline{\phi}(\varepsilon) \leqq \mathcal{N}(\varepsilon) \leqq \bar{\phi}(\varepsilon) \leqq B_2\varepsilon^{-D}$ for $0 < \varepsilon \leqq \varepsilon_0$.

Case 2. $\gamma = \sum_{i=1}^{N} |a_i| < 1$.

As in Case 1, select $k_2 > 0$ so that $\mathcal{N}^*(\varepsilon) \leqq \bar{\phi}(\varepsilon) = (\beta_2/(1-\gamma))\varepsilon^{-1} + k_2\varepsilon^{-D}$ for $0 < \varepsilon \leqq 1$. Since $D < 1$ and $\beta_2/(1-\gamma) > 0$ we obtain $\mathcal{N}(\varepsilon) \leqq B_2\varepsilon^{-1}$ for some $B_2 > 0$.

As G is the graph of a continuous function on $[0, 1]$ we have $\mathcal{N}(\varepsilon) \geqq 1/\varepsilon$ and thus $\dim(G) = 1$.

Case 3. $\gamma = \sum |a_i| = 1$.

Let $\bar{\phi}(\varepsilon) = (\beta_2/\sum_{i=1}^{N} |a_i| \ln(b_i))(\ln(\varepsilon)/\varepsilon) + k_2/\varepsilon$ and note that

$$\bar{\phi}(\varepsilon) = \sum_{i=1}^{N} \frac{|a_i|}{b_i}\bar{\phi}\left(\frac{\varepsilon}{b_i}\right) + \frac{\beta_2}{\varepsilon}$$

for $\varepsilon > 0$ and any $k_2 \in \mathbb{R}$. As in Cases 1 and 2, select $k_2 > 0$ so that $\mathcal{N}(\varepsilon) \leqq \bar{\phi}(\varepsilon)$ for $0 < \varepsilon \leqq 1$, which shows that $\dim(G) = 1$.

Case 4. $\{(t_j, x_j): j = 0, \cdots, N\}$ is collinear.

G is a line segment and thus $\dim(G) = 1$. □

4.2. Dimension of projections. Let us now calculate the dimension of $g = \text{graph}(f)$ under the assumption that (3.2)(a) is satisfied and that $J = [0, 1]$.

THEOREM 5. *Let $f: I \to \mathbb{R}$ be defined as in § 1 above and let g be its graph. Suppose that A satisfies* (3.2). *Then* $\dim(g)$ *is the unique positive solution d of*

$$\sum_{i=1}^{N} s_i b_i^{d-1} = 1 \tag{4.4}$$

where $b_i = t_i - t_{i-1}$, for all i.

Proof. We will use the notation introduced in the proof of Theorem 4. Let $\{\tau_l\}_{l=0}^{m}$ be an ε-partition with associated ε-column cover $\mathscr{C}$. Assume that $\mathscr{C}$ is also a minimal cover of g. Choose $k \in \{1, \cdots, m\}$ and consider the collection C_k of all $\varepsilon \times \varepsilon$ squares from $\mathscr{C}$ that lie above $[\tau_{k-1}, \tau_k]$. The continuity of f implies that C_k is a rectangle of width ε and height h_k. Let $\Pi = \{\Pi_\theta: \theta \in [0, 2\pi)\}$ be a plane bundle whose axis is the t-axis. Let $\Pi_\theta \in \Pi$ and let $h_k(\theta)$ be the height of the projection of $g|_{[\tau_{k-1}, \tau_k]}$ onto Π_θ. Denote by $\bar{\delta}_k(\underline{\delta}_k)$ the maximum (minimum) of $h_k(\theta)$ over $\theta \in [0, 2\pi)$.

Define $\bar{\mathcal{N}} = \sum_k \bar{\delta}_k/\varepsilon$ and $\underline{\mathcal{N}} = \sum_k \underline{\delta}_k/\varepsilon$. Then

$$\underline{\mathcal{N}}(\varepsilon) \leqq \mathcal{N}^*(\varepsilon) \leqq \bar{\mathcal{N}}(\varepsilon).$$

Following the argument given in the proof of Theorem 4 we obtain functional inequalities for $\underline{\mathcal{N}}(\varepsilon)$ and $\bar{\mathcal{N}}(\varepsilon)$. (Note that $\{(\tau_j, x_j): j \in \{0, \cdots, N\}\}$ is collinear $\Leftrightarrow$ $\sum_{i=1}^{N} s_i = 1$.) These functional inequalities then imply that

$$\underline{\mathcal{N}}(\varepsilon) \geqq A\varepsilon^{-d} \quad \text{and} \quad \bar{\mathcal{N}}(\varepsilon) \leqq B\varepsilon^{-d}$$

for positive constants A and B and where d is the unique real solution of $\sum_{i=1}^{N} s_i b_i^{d-1} = 1$. □

Example 8. Let $K = [0, 1] \times [0, 1]$ and let g be the graph of f, which is the projection of the attractor G of the IFS $(J \times K, W_i, i \in \{1, \cdots, 3\})$, where the maps $W_i: J \times K$ are defined by

$$W_i(t, X) = \begin{pmatrix} b_i & 0 & 0 \\ 0 & a_i & -c_i \\ 0 & c_i & a_i \end{pmatrix} (t, X) + \begin{pmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{pmatrix}$$

and the constants a_i, b_i, c_i, α_i, β_i, and γ_i are determined by (2.7) with $\{(0, 0, 0), (\frac{3}{10}, \frac{7}{10}, \frac{2}{5}), (\frac{7}{10}, \frac{2}{5}, \frac{3}{10}), (1, 1, 0)\}$ as the set of interpolation points. Then $b_1 = \frac{3}{10}$, $b_2 = \frac{2}{5}$, $b_3 = \frac{3}{10}$, and the scaling factors are $s_i = \sqrt{65}/10$, $s_2 = \sqrt{10}/10$, and $s_3 = \sqrt{45}/10$. The dimension d of g is then approximately $d \approx 1.5075$. Figure 15 shows the projections of G.

If in addition we now assume that the attractor A of the IFS $(K, k_i, i = 1, \cdots, N)$ also satisfies (3.2b), then we can derive the following relation between $\dim(g)$ and $\dim(K) = n$.

COROLLARY 2. Under the hypotheses of Theorem 5 and the assumption that (3.2b) holds, we have that

$$1 \leqq \dim(g) \leqq 2 - \frac{1}{n}. \tag{4.5}$$

Proof. The statement follows readily from the Cauchy-Schwartz inequality applied to (4.4) and the fact that $\Sigma\, s_i \geqq 1$ and $\Sigma\, s_i^n \leqq 1$ (the first inequality reflects the connectedness of A, and the second, the fact that $\dim(A) \leqq n$). □

Example 9. For the attractor G in Example 2 we have $s_1 = s_2 = s_3 = \frac{1}{2}$ and $b_1 = b_2 = b_3 = \frac{1}{3}$. Hence $d = 2 - \log 2/\log 3 \approx 1.3691 < 1.5$.

Example 10. For the attractor of Example 4 we have $s_1 = s_2 = 1/\sqrt{2}$ and $b_1 = b_2 = \frac{1}{2}$. Thus $d = 1 + \log_2 \sqrt{2} = 1.5$.

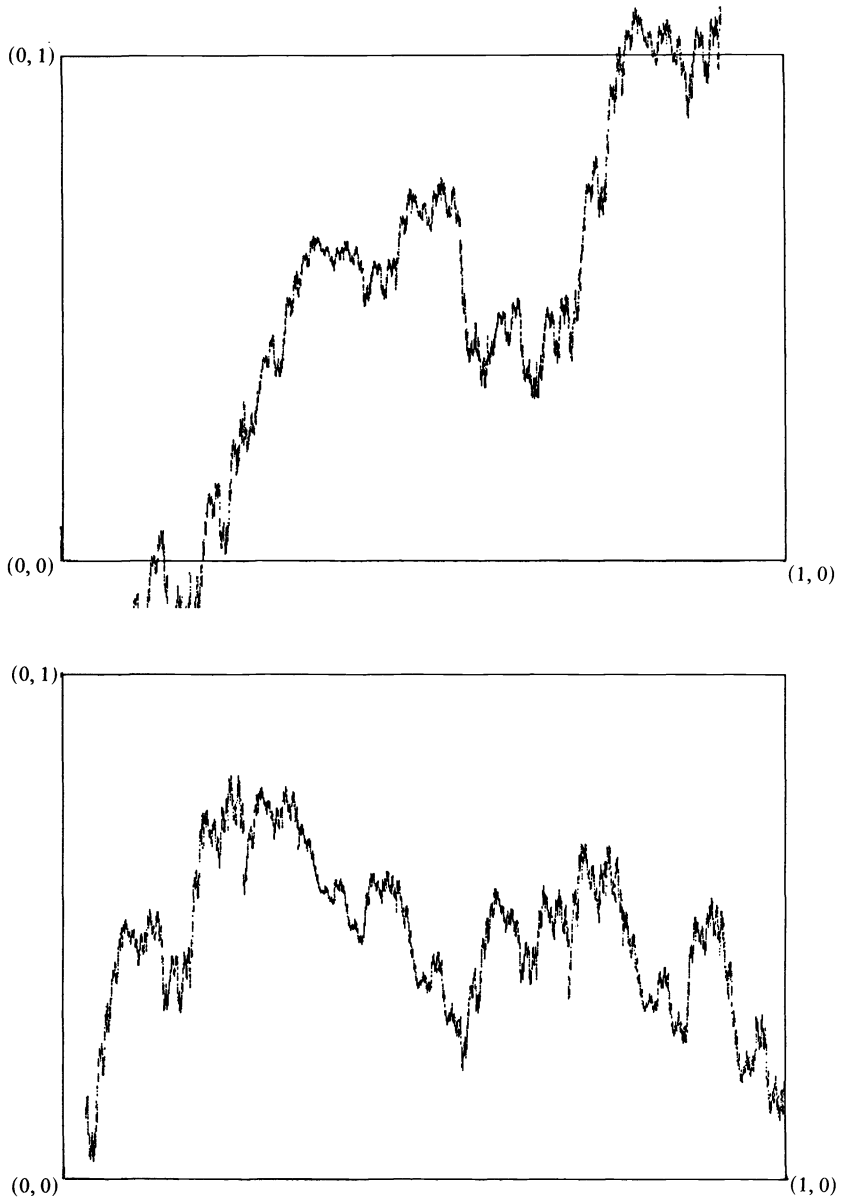


FIG. 15. *The projections of the attractor G for the* IFS *in Example* 8.

Example 11. Let $J=[0,1]$ and $K=[0,1]\times[0,1]\times[0,1]$. Define maps $W_i: J\times K$ by

$$W_i(t,X)=\begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & a_i & c_i & d_i \\ 0 & e_i & f_i & g_i \\ 0 & h_i & k_i & l_i \end{pmatrix}(t,X)+\begin{pmatrix} (i-1)/4 \\ \alpha_i \\ \beta_i \\ \gamma_i \end{pmatrix}$$

for all $i\in\{1,\cdots,4\}$, where the constants a_i, c_i, d_i, e_i, f_i, g_i, h_i, k_i, l_i, α_i, β_i, and γ_i are determined by (2.7) with $\{(t_j,X_j): j\in\{0,\cdots,4\}\}=\{(0,0,0,0),\ (\frac{1}{4},\frac{1}{4},\sqrt{3}/8,\sqrt{6}/6),$

$(\frac{1}{2}, \frac{1}{2}, \sqrt{3}/4, \sqrt{6}/6), (\frac{3}{4}, \frac{3}{4}, \sqrt{3}/4, 0), (1, 1, 0, 0)\}$. Then $s_1 = s_2 = s_3 = s_4 = \frac{1}{2}$ and $b_1 = b_2 = b_3 = b_4 = \frac{1}{4}$. Hence $d = 1.5 < 2 - \frac{1}{3} = \frac{8}{3}$.

The following example due to M. Berger shows that if (3.2a) does not hold, then (4.5) need not be true.

Example 12. Let $J = [0, 1]$ and $K = [0, 1] \times [0, 1]$ and define maps $W_i: J \times K$ by

$$W_i(t, X) = \begin{pmatrix} b_i & 0 & 0 \\ 0 & b_i & 0 \\ 0 & \Delta y_i & a_i \end{pmatrix} (t, X) + \begin{pmatrix} x_{i-1} \\ X_{i-1} \end{pmatrix}$$

where $X_i = \binom{x_i}{y_i}$, $b_i = x_i - x_{i-1}$, $\Delta y_i = y_i - y_{i-1}$, and $|a_i| < 1$, for all $i \in \{1, \cdots, N\}$, for some positive integer $N > 1$. Suppose further that (4.1) holds.

From Theorem 4 we know then that the attractor A of the IFS $(J \times K, W_i, i \in \{1, \cdots, N\})$ is the graph G of a continuous function (see also Example 1) and that $d = \dim(G)$ satisfies $\sum_{i=1}^{N} |a_i| b_i^{d-1} = 1$. For all $\varepsilon > 0$ we can choose a_i and b_i such that $2 - d < \varepsilon$. Let g be the projection of G onto the ty-plane. Then g is identical to G since the scalings along the t-axis are the same as along the x-axis. Hence $\dim(g) = \dim(G) = d$, and we see that (4.5) need not hold.

Acknowledgments. We thank Dr. Marc Berger for providing Example 12 and Norman Fickel for correcting an error in the proof of Theorem 4.

REFERENCES

[B] M. F. BARNSLEY, *Fractal functions and interpolation*, Constr. Approx., 2 (1986), pp. 303-329.

[BD] M. F. BARNSLEY AND S. DEMKO, *Iterated function systems and the global construction of fractals*, Proc. Roy. Soc. London Ser. A, 399 (1985), pp. 243-275.

[BH] M. F. BARNSLEY AND A. N. HARRINGTON, *The calculus of fractal interpolation functions*, J. Approx. Theory, 57 (1989), pp. 14-34.

[DHN] S. DEMKO, L. HODGES, AND B. NAYLOR, *Construction of Fractal Objects with Iterated Function Systems*, Computer Graphics, 19 (1985), pp. 271-278.

[HM] D. P. HARDIN AND P. R. MASSOPUST, *The capacity of a class of fractal functions*, Commun. Math. Phys., 105 (1986), pp. 455-460.

[H] J. HUTCHINSON, *Fractals and self-similarity*, Indiana Univ. J. Math., 30 (1981), pp. 731-747.

[M] P. R. MASSOPUST, *Hidden variable fractal interpolation functions*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1986.

[Mo] P. A. P. MORAN, *Additive functions on intervals and Hausdorff measure*, Proc. Cambridge Philos. Soc., 42 (1946), pp. 15-23.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1243-1254, September 1989

013

THE DIMENSION SPECTRUM OF THE MAXIMAL MEASURE*

ARTUR O. LOPES†

Abstract. A variety of complicated fractal objects and strange sets appears in nonlinear physics. In diffusion-limited aggregation, the probability of a random walker landing next to a given site of the aggregate is of interest. In percolation, the distribution of voltages across different elements in a random-resistor network (see [T. Halsey et al., *Phys. Rev. A* (3), 33 (1986), pp. 1141-1151]) may be of interest. These examples can be better analyzed by dividing certain objects in pieces labeled by indexes, but that leads to working with fractal sets and the notion of dimension [Halsey et al. (1986)].

The dimension spectrum of a system has been introduced and measured experimentally, and a substantial literature in physics addresses this topic. In several important cases, rigorous proofs of the results presented in [Halsey et al. (1986)] have been established.

Here, rigorous mathematical proofs of some results in this theory are given, specifically for the maximal entropy measure of a hyperbolic rational map in the complex plane. In this case the fractal object is the Julia set (see [H. Brolin, *Ark. Mat.*, 6 (1966), pp. 103-114], [A. Freire, A. Lopes, and R. Mañé, *Bol. Soc. Brasil Mat.*, 14 (1983), pp. 45-62]), which has been extensively studied in the physics literature.

Key words. Hausdorff dimension, entropy, maximal measure, rational maps, pressure, spectrum of dimension, large deviation

AMS(MOS) subject classification. 58F11

0. Introduction. In recent years the role of the concept of dimension has been investigated by several authors in trying to understand nonconservative dynamical systems.

The possibility of an infinite number of generalized dimensions of fractals appears in a natural way in the context of relevent physical problems of critical phenomena. This topic is particularly active in the physics literature. Such problems appear in the configuration of Ising models, percolation clusters, and fully developed turbulence. In general, we can describe such models by dividing the object into pieces and rescaling. In this situation we very often obtain several different values of dimension.

We are interested in developing the thermodynamic formalism for chaotic repellers obtained from hyperbolic rational maps in the complex plane and its relation to the spectrum of dimensions.

The same problem for attractors has been investigated in [9]. In general, an attractor can have an arbitrarily fine-scaled interwoven structure of hot and cold spots (high and low probability densities). By hot and cold spots we mean points on the attractor for which the frequency of visitation to the region for typical orbits is either much greater than average (a hot spot) or much less than average (a cold spot). In these several different points we can have different local values of dimension, and the aim of this theory is to understand the situation globally.

Now we will explain more carefully the situation we are going to consider. We will analyze the dimension spectrum of the maximal measure (sometimes called the balanced measure) [1], [8], [17] of a hyperbolic rational map f on the complex plane

$$f(z)=\frac{P(z)}{Q(z)}$$

where P and Q are complex polynomials.

* Received by the editors January 25, 1988; accepted for publication (in revised form) November 22, 1988. This work was partially supported by Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (Brazil) and by the U.S. Air Force Office of Scientific Research.

† Instituto de Matematica, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. Present address, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

The dimension spectrum of a system was introduced and measured experimentally by Halsey et al. [10], Hentschel and Procaccia [11], and Jensen et al. [12]. See also [2], [5], [9], [26], and [27] for analyses of important cases.

In [2] and [27] the theory is applied to several different systems, among them cookie-cutter maps, and it is related to the measure of maximal entropy. In [5] critical mappings of the circle with golden rotation number are considered.

We will use thermodynamic formalism as in [27] and also classical large deviation theory as in [5] to obtain our result for the maximal measure of a hyperbolic rational map.

For each complex number z and positive real number ξ, denote by $B(z, \xi)$ the ball of center z and radius ξ in the usual norm of $\mathbf{R}^2$.

We will say that a certain measure v has exponent α on z if

$$v(B(z, \xi)) \approx \xi^{\alpha}$$

for ξ small enough. (Here $\approx$ means $\lim_{\xi \to 0} (\log v(B(z, \xi))/\log \xi) = \alpha$.)

We will also say that z scales with exponent α.

Given a measure v, one of the main goals of the Dimension Spectrum Theory is to understand the set of points that scales with exponent α.

For α fixed, the structure of such a set of points can be very complicated, and this set can also have v-measure zero and two-dimensional Lebesgue measure zero. The Hausdorff dimension gives more detailed information on how small the sets of the plane with two-dimensional Lebesgue measure zero are. When the Hausdorff dimension of a set is a noninteger number, we say that this set is a fractal. It is natural to ask, in terms of Hausdorff dimension, how small these sets are with respect to the variable α.

Experimental results in [10] and [12] have suggested that the Hausdorff dimension of such sets is a differentiable function of α, in the case of a certain measure of critical mappings of the circle with rotation number equal to the golden-mean.

We point out, as has been done in [5], that without some restrictions on the measure v, nothing interesting can be said about the problem.

A given probability v is called invariant for a map f if

$$v(f^{-1}(E)) = v(E)$$

for any set E, where the probability is defined.

If we are working in the context of statistical physics with problems in the one-dimensional $\mathbf{Z}$ lattice, and in each position we have two possibilities of spin, let us say $+$ and $-$, then the natural space to consider is the Bernoulli model $\{+, -\}^{\mathbf{Z}}$. As we do not have any reason to consider a distinguished position for the value zero in our lattice, then in our problem we will consider only probabilities that are invariant by the shift map (see [3] and [29] for more references). This is a simple motivation for considering invariant probabilities in general problems.

In cases where f is a rational map, the support of any invariant probability is the Julia set (see [4], [6], [8], [21] for definitions). In almost all the cases this set is of fractal dimension [6], [14]. There are no smooth invariant measures to consider in this situation.

Consider, for example, the map $f_{\xi} = z^2 + \xi$ when ξ is small. In this case the Julia set is a nowhere-differentiable Jordan curve for $\xi \neq 0$. In fact, the Julia set is a fractal Jordan curve for ξ inside the main cardioid of the Mandelbrot set (and $\xi \neq 0$) [6].

The Julia set can also be a Cantor set or even a combination of parts that are locally disconnected and locally connected. The Julia set can even be the all complex plane for some nonhyperbolic rational maps. In the case of hyperbolic rational maps anyway, the Julia set always has two-dimensional Lebesgue measure zero.

There is an important conjecture that claims that the hyperbolic rational maps are dense in the set of rational maps (see [21]).

In [8] and [17] it has been shown that among all invariant probabilities, there exists a special one that obtains the maximal value of the entropy (see [19], [33] for exact definitions). We will call this probability the maximal measure.

The entropy of an invariant probability is a measure of the degree of randomness of the system given by the action of the map f and the invariant probability we are considering. In this case the maximal measure is the more chaotic one.

Following the principle that in the absence of external thermal sources nature tends to maximize entropy, we can see the maximal measure as some kind of Gibbs state. If we must take into account external sources, we are then led to consider maximal pressure probabilities (see [3], [29] for interesting considerations about this). In § 1, for some other reasons, we will have to consider maximal pressure probabilities.

We will denote by u the maximal measure for a hyperbolic rational map. We point out that, in [8] and [17], the results are for general rational maps, and hyperbolicity is not assumed.

Here we will develop all the theory to show the following theorem.

THEOREM. *Consider u the maximal measure of a hyperbolic rational map; then the Hausdorff dimension of the set of points that scale with exponent α is a real analytic function of the variable α.*

We will relate these concepts of scaling exponents with the pressure, the Legendre transform of the pressure, entropy, and large deviation. In fact, one of the main ingredients of the proof is the close relation of pressure and free-energy (see § 1 for definitions). This relationship is explored in a more general context in [16].

The analogous claim for nonhyperbolic rational maps is not always true. In [26] an example of a quadratic polynomial is shown such that there exists a point α where there is no differentiability. In this situation we can say, using an analogy with statistical physics, that phase transition exists.

The theorem stated here can also be seen as a statement concerning the nonexistence of phase transitions for the maximal entropy measure of a hyperbolic rational map.

A natural question to ask is, why do large deviation techniques appear in the understanding of the problem? The reason is that there exists a certain δ such that for a u-almost-everywhere point z in the Julia set, the point z scales with exponent δ (see [22], [20]). This follows basically from properties related to the Birkhoff Ergodic Theorem [19]. If we want to consider a certain fixed α different from the above-mentioned δ and look for the set of points z that scales with exponent α, then we are in part not covered by the Birkhoff Ergodic Theorem. The above-mentioned theorem is a result on mean values and, therefore, in considering deviations of the mean, we must use large deviation techniques. We refer the reader to Ellis [7] for references concerning large deviation. We can find the general theory of ergodic theory and thermodynamic formalism in Walters [33], Mañé [19], and Ruelle [29], [30].

Several results are known for the maximal measure [1], [8], [13]–[15], [17], [18], [20], [22]. In particular, the moments of this measure can be obtained by a three-term relation from the coefficients of the rational map (see [1], [13]). The three-term relation is a consequence of the functional equation that the complex potential generated by

the maximal measure satisfies around infinity [13]. This functional equation is known as the Bochner equation in the case of polynomial maps [1]. In the case where the rational map is not a polynomial, the functional equation has another form (see [13], [14]). There are several connections of such results with Classical Potential Theory [31], [4], [13]. In particular, [4] and [13] show that this maximal measure is the charge distribution in the Julia set if and only if f is a polynomial.

The degree of the rational map f will be denoted by d. We will also denote by J the Julia set. The entropy of the maximal measure is $\log d$ [8], [17].

We will show here that, in fact, the set of points that scale with exponent α can be considered as the support of another measure (different from u, and we will not lose dimensionality with this procedure (see the proof of Theorem 4)). We refer the reader to [9]–[12] and [26] for some applications of the spectrum of dimension theory to statistical mechanics.

It is worthwhile mentioning the following heuristic analogy. In problems of physics, when we can apply renormalization techniques, in general, it is because we have some good self-similar properties. We can take a partition of the object we want to consider, and from this partition, by some well-defined procedure, we can obtain another with some additional coarse information. Now, the procedure is repeated with the new partition. If we have some good self-similar properties, we can expect to have with this procedure microscopic information from the macroscopic information. In this case, scaling properties appear in a natural way. The spectrum of dimension techniques are suitable for application in this situation. Perhaps one reason this theory works well for a rational map f is because we can think of the inverse branches of f as a natural way to obtain new partitions. Because these inverse branches are holomorphic, we have good self-similarity properties that come from the conformality and from the Koebe Distortion Theorem (see [8]).

Here is the structure of this paper. In § 1 we will introduce the main properties of ergodic theory and large deviations that we will use. In § 2 we will present the main theorem and give an outline of the proof. In § 3 we will give the formal proof of the main theorem.

1. Ergodic theory and large deviation. Let $M(f)$ be the set of invariant probabilities for f, that is, the set of measures v such that $v(f^{-1}(A)) = v(A)$ for any set A in the Borel sigma-field of $\mathbf{R}^2$. The support of all these measures is J.

DEFINITION 1. For a Hölder continuous $g: J \to \mathbf{R}$ and $v \in M(f)$, we will define the pressure of v with respect to g by

$$h(v) + \int g(z)\, dv(z)$$

where $h(v)$ denotes the entropy of v. We will denote such an expression by $P(v, g)$.

DEFINITION 2. We will call $P(g) = \sup \{P(v, g) | v \in M(f)\}$ the topological pressure of the function g.

In the case where f is hyperbolic, there exists a unique measure that attains such supremum. This measure is ergodic. These measures are sometimes called Gibbs measures [3], [29], [30]. There exist examples of C^k maps such that this supremum is not attained (see references in [19], [33]).

DEFINITION 3. In the case where there exists a unique probability in $M(f)$, denoted by $v(g)$, such that $P(g) = h(v(g)) + \int g(z) d(v(g))(z)$ we will call this measure the maximal pressure measure for $g: J \to \mathbf{R}$. When f is hyperbolic, this is always the case [3], [28].

DEFINITION 4. For g constant and equal to zero, the maximal pressure measure is called the maximal measure.

Let z_0 be a point in the Riemann sphere, and for each $n \in \mathbf{N}$, let us denote by $z(n, i, z_0)$, $i \in \{1, 2, \cdots, d^n\}$ the d^n-solutions (with multiplicity) of the equation

$$f^n(z) = z_0.$$

We denote the delta Dirac measure on z by $\delta(z)$.

Let $u(n, z_0)$ be the probability

$$d^{-n} \sum_{i=1}^{d^n} \delta(z(n, i, z_0)).$$

In [8] and [17], it has been shown that for any z_0 (but at most two exceptional points), and independent of z_0, there exists the weak limit

$$\lim_{n\to\infty} u(n, z_0) = u,$$

and the measure u is the maximal measure of the rational map f. Hyperbolicity is not assumed to obtain this result. Also, u is ergodic and has entropy $\log d$. We will denote z_i^n the $z(n, i, z_0)$ for a certain fixed z_0.

DEFINITION 5. For any real $t \in \mathbf{R}$ we will denote $P(t) = P(g)$, when $g(z) = -t \log |f'(z)|$.

From [23] and [24] it is known that $P(t)$ is convex and real analytic in the variable t when f is hyperbolic.

DEFINITION 6. For a given probability v we will call the Hausdorff dimension of the measure v, denoted by $\mathrm{HD}(v)$, the value $\inf\{\mathrm{HD}(A) | v(A) = 1, A \text{ a Borel set in } J\}$. Here $\mathrm{HD}(A)$ is the Hausdorff dimension of the set A.

DEFINITION 7. For any real $t \in \mathbf{R}$ we will denote $u(t)$ as the maximal pressure measure for $g(z) = -t \log |f'(z)|$.

It also follows from [20], [22], and [24] that if f is hyperbolic, then

$$P'(t) = -\int \log |f'(z)| d(u(t))(z) = -h(u(t)) \cdot (\mathrm{HD}(u(t)))^{-1}.$$

THEOREM 1 [20]. *Let f be a rational map and let $v \in M(f)$ be an ergodic probability; then there exists a Borel set A such that $v(A) = 1$, and for all $z \in A$,*

$$\lim_{r\to 0} \frac{\log v(B(z, r))}{\log r} = h(v)\left(\int \log |f'(z)| \, dv(z)\right)^{-1} = \mathrm{HD}(v)$$

where $B(z, r)$ denotes the disk of radius r and center at z.

THEOREM 2 [28]. *If f is a hyperbolic rational map, then*

$$P(t) = \lim_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^{d^n} |(f^n)'(z_i^n)|^{-t}.$$

In § 2 we will explain why we need the pressure in this formulation.

Let $W = \{W_n : n = 1, 2, 3, \cdots\}$ be a sequence of random variables that are defined on probability spaces $\{(\tau_n, \mathcal{F}_n, P_n), n = 1, 2, \cdots\}$ and that take values in $\mathbf{R}$, where τ_n is a set, $\mathcal{F}_n$ a σ-field, and P_n a probability.

Here we will consider $\tau_n = J$, and $\mathcal{F}_n$ the Borel σ-field on J, $n \in \mathbf{N}$.

DEFINITION 8. For each $n \in \mathbf{N}$ define

$$c_n(t) = n^{-1} \log E_n\{\exp tW_n\}$$

where E_n is the expected value with respect to P_n.

We will consider in this case the weak topology in the space of signed-measures in J.

The following hypotheses are assumed to hold:

(a) Each function $c_n(t)$ is finite for all $t \in \mathbf{R}$.

(b) $c(t) = \lim_{n\to\infty} c_n(t)$ exists and is finite for all $t \in \mathbf{R}$.

(c) $c(t)$ is differentiable as a function of $t \in \mathbf{R}$.

THEOREM 3 [7]. *Assume hypotheses* (a), (b), *and* (c) *hold, and denote for each compact Borel set K in E*

$$Q_n(K) = P_n\{z \in J | n^{-1} W_n \in K\} \quad \text{and} \quad I(z) = \sup_{t\in\mathbf{R}} \{zt - c(t)\}, \quad z \in \mathbf{R}.$$

Then the following conclusion holds:

$$\lim_{n\to\infty} n^{-1} \log Q_n(K) = -\inf_{z\in K} \{I(z)\}.$$

DEFINITION 9. The function $c(t)$ is called the free energy of W_n.

DEFINITION 10. The function $I(z)$ is called the deviation function of the process [7]. In fact it is the Legendre transform of $c(t)$. The function $I(z)$ contains information about the deviations of the mean of the process.

2. An outline of the proof of the main theorem. Some ideas presented here were adapted from ideas in [5] and [27].

It follows from [8] that in the hyperbolic case, for any $z \in J$ (this is not a u-almost-everywhere statement), there exists

$$\lim_{\varepsilon\to 0} \lim_{n\to\infty} n^{-1} \log u(B(z, n, \varepsilon)) = -\log d$$

where $B(z, n, \varepsilon) = \{y \in J | j \in \{0, 1, \cdots, n-1\}, |f^j(y) - f^j(z)| < \varepsilon\}$.

It is also true in the hyperbolic case that the diameter $d(z, \nu, \varepsilon)$ of $B(z, n, \varepsilon)$ is of the order $|(f^n)'(z)|^{-1}$, for any $z \in J$, for n large and ε small [8], [22].

Therefore, if we ask whether z is such that $u(B(z, \xi)) \approx \xi^\alpha$, it is natural to consider the above definition.

DEFINITION 11. Let $J(\alpha)$ be the set of points $z \in J(f)$ such that there exists the limit

$$\lim_{n\to\infty} n^{-1} \log |(f^n)'(z)|^{-\alpha} = -\log d.$$

In this case, $u(B(z, n, \varepsilon))$ is of order $d(z, n, \varepsilon)^\alpha$. When we use arguments of [20] it follows that this is equivalent to requiring that z satisfies

$$\lim_{\xi\to 0} \frac{\log u(B(z, \xi))}{\log \xi} = \alpha.$$

DEFINITION 12. Let $\mathfrak{f}(\alpha)$ be the Hausdorff dimension of the set $J(\alpha)$.

THEOREM 4. *Suppose f is a hyperbolic rational map and u the maximal measure. Then for a given α, there exists a unique $t \in \mathbf{R}$ such that $P'(t) = -\log d/\alpha$, and $\mathfrak{f}(\alpha) = \mathrm{HD}(u(t))$, where $u(t)$ is the maximal pressure measure for $-t \log |f'(z)|$. The function $\mathfrak{f}$ is real analytic on α.*

We will now give an outline of the proof of Theorem 4 as we mentioned in the Introduction.

The proof is divided in two parts; this is very characteristic of large deviation results [7]. We must deal with the lower bound and the upper bound in separate cases.

In the first part, we show $f(\alpha) \geqq \mathrm{HD}(u(t))$, where t satisfies a Legendre condition of the form $P'(t) = -\log d/\alpha$. This part can be seen as an application of the formula

$\mathrm{HD}(v) = h(v)/\int \log |f'(z)| dv(z)$ (that is true for any invariant measure v [20], [22]) and the Manning-McCluskey picture, which means in our case that $\mathfrak{f}(\alpha)$ is the Legendre transform of the pressure [24].

The pressure contains information about $u(t)$ in the form

$$P'(t) = -\int \log |f'(z)| d(u(t))(z).$$

This information is about the Lyapunov number of $u(t)$. Using this information we obtain a set with dimension $\mathrm{HD}(u(t))$ such that for any point z on it, the measure u scales with exponent α in z. This set is the support of the measure $u(t)$. In this way we show $\mathfrak{f}(\alpha) \geqq \mathrm{HD}(u(t))$.

Now, in the second part, it is more difficult to show that $\mathfrak{f}(\alpha) \leqq \mathrm{HD}(u(t))$.

We will try to give a heuristic idea of the proof, even under the risk of oversimplifying some more difficult and subtle parts of the demonstration. First, to have a geometrical picture of the problem, consider for simplification $f(z) = z^2 + \xi z$, when ξ is small. In this case $d = 2$. Note that zero is a fixed point of f. The main ideas of the proof are presented in this simplified case. The Julia set in this situation is a nowhere-differentiable Jordan curve. This curve is very close to the unitary circle and the dynamics of f is very similar to that of z^2 on the unitary circle (they are in fact topologically conjugated). Now consider a nonself-intersecting curve γ_1^0, from zero to ∞, cutting the Julia set in the unique fixed point in this set. Taking pre-images of this curve, we obtain the new curves γ_1^1 and γ_2^1. The Julia set without these two curves has two connected components denoted by A_1^1 and A_2^1, each one with u-measure $d^{-1} = \frac{1}{2}$.

Now consider γ_1^2, γ_2^2, γ_3^2, and γ_4^2, the pre-images of the curves γ_1^1 and γ_2^1. Now the Julia set without these four curves has four connected components denoted by A_1^2, A_2^2, A_3^2, and A_4^2. Each of these components has measure $d^{-2} = 2^{-2}$. Repeating the procedure inductively, we obtain at level n, a total of $d^n = 2^n$ curves $\gamma_1^n, \gamma_2^n, \cdots, \gamma_{2^n}^n$. The Julia set without these 2^n curves has 2^n connected components denoted by $A_1^n, A_2^n, \cdots, A_{2^n}^n$, each one with u-measure $2^{-n} = d^{-n}$. If we select an initial point z_0 not in γ_1^0, then we can suppose that in each A_i^n, $i \in \{1, 2, \cdots, 2^n\}$ there exists one and only one $z(n, i, z_0)$ (see the notation in § 1).

Now we look at level n, which has the elements of the partition $A_1^n, A_2^n, \cdots, A_{2^n}^n$ that contains elements z such that $|f^{n'}(z)|^{-\alpha}$ is of order d^{-n}. By the Koebe Distortion Theorem (see [8]) we conclude (in fact, we have to consider subsets of the A_i^n, $i \in \{1, \cdots, 2^n\}$, but we do not want to be too technical here in § 2) that if A_i^n contains a z such as the one above, then the $z(n, i, z_0)$ contained in A_i^n also has this property.

Note that from the Birkhoff Ergodic Theorem (concerning mean values) and the Shannon-McMillan-Breiman Theorem (about entropy of partitions), almost all the $z(n, i, z_0)$ should satisfy

$$|f^{n'}(z(n, i, z_0))|^{-\mathrm{HD}(u)}$$

and be of order d^{-n}. This is a simplified way to look at the formula

$$\mathrm{HD}(u) = \frac{h(u)}{\int \log |f'(z)| \, du(z)}.$$

Therefore, the large deviation here appears to give information on how many elements $z(n, i, z_0)$, $i \in \{1, 2, \cdots, 2^n\}$ deviate from the mean and satisfy that $|(f^n)'(z(n, i, z_0))|^{-\alpha}$ is of order d^{-n}.

Here it becomes clear why we must consider the pressure $P(t)$ in the formulation given by Theorem 2. We must consider the random variable given by $-\log|f^{n'}(z)|$ in the pre-orbits of z_0 at level n. At this moment the close relation of $c(t)$ and $P(t)$, which we will explain in § 3, is essential.

The diameter of each element A_i^n of the partition is of order $|f^{n'}(z(n, i, z_0))|^{-1}$, where $z(n, i, z_0)$ is the only pre-image of z_0 at level n in A_i^n.

From the considerations above, we can cover the set of points that scale with exponent α with a controlled number of elements of the partition, and we also have control of the diameter of the elements of the partition that we are using to cover the set $J(\alpha)$. This partition can be obtained with a diameter as small as we want. The value $\mathrm{HD}(u(t))$ (it appears here as information that comes from a Legendre transform) is exactly the value that we must consider for the Hausdorff measure to be finite. In this way we prove finally that $\mathcal{f}(\alpha) \leqq \mathrm{HD}(u(t))$.

The above explanation is not exactly as the proof will be done, but it gives a good idea of the main ingredients of the demonstration.

3. Proof of the main theorem. Here we will show the proof of the following theorem.

THEOREM. *Suppose f is a hyperbolic rational map and u is the measure of maximal entropy. Then for a given α, there exists a unique $t \in \mathbf{R}$ such that $P'(t) = -\log d/\alpha$ and $\mathcal{f}(\alpha) = \mathrm{HD}(u(t))$, where $u(t)$ is the maximal pressure measure for $-t\log|f'(z)|$. The function $\mathcal{f}$ is real analytic in the variable α.*

Proof. (a) $\mathcal{f}(\alpha) \geqq HD(u(t))$.

For a given α, from the convexity and analyticity of $P(t)$ (see [28]), we have that there exists a unique t such that $P'(t) = -\log d/\alpha$. For this value of t, consider $u(t)$ the maximal pressure measure for the function $g = -t\log|f'|$. From the ergodic theorem we have that for a set A such that $u(t)(A) = 1$, for all $z \in A$,

$$\lim_{n\to\infty} \frac{1}{n} \log|(f^n)'(z)|^{-1} = -\int \log|f'(z)| d(u(t))(z) = P'(t) = -\frac{\log d}{\alpha}.$$

Therefore

$$\lim_{n\to\infty} n^{-1} \log|f^{n'}(z)|^{-\alpha} = -\log d$$

and $A \subset J(\alpha)$.

As the Hausdorff dimension of $u(t)$ is infimum of the Hausdorff dimension of all sets of measure zero, we have $\mathcal{f}(\alpha) \geqq \mathrm{HD}(u(t))$.

(b) $\mathcal{f}(\alpha) \leqq \mathrm{HD}(u(t))$.

Now we will use a large deviation property as introduced in § 1.

Consider for each $n \in \mathbf{N}$ the measure $u(n, z_0)$ as defined in § 1.

We will denote z_i^n the $z(n, i, z_0)$ to make the notation simpler. To apply Theorem 3, consider $\tau_n = J$, $\mathcal{F}_n =$ Borel σ-field on J, and $P_n = u(n, z_0)$. Consider also the random variable $W_n = -\log|f^{n'}(z)|$.

From Theorem 3 we have that

$$\lim_{n\to\infty} n^{-1} \log \sum_{i=1}^{d^n} |f^{n'}(z_i^n)|^{-t} = P(t)$$

$$= \sup_{v \in M(f)} \left\{ h(v) - t \int \log|f'(z)|\, dv(z) \right\}.$$

Therefore

$$c(t) = \lim_{n\to\infty} n^{-1} \log E_n\{\exp t\, W_n\}$$

$$= \lim_{n\to\infty} n^{-1} \log d^{-n}\left(\sum_{i=1}^{d^n} |f^{n'}(z_i^n)|^{-t}\right) = P(t) - \log d.$$

This relation of pressure and free energy is essential for the rest of the proof.

From the differentiability with respect to t of $P(t)$ [16], [19], we have that for any $\beta \in \mathbf{R}$ and $\xi > 0$

$$\lim_{n\to\infty} n^{-1} \log P_n\{n^{-1} W_n \in (\beta - \xi, \beta + \xi)\}$$

is almost equal to $-I(\beta)$, where $I(\beta) = \sup_{s\in R}\{s\beta - c(s)\}$.

Therefore for $\beta = -\log d/\alpha$, we have $I(\beta) = t\beta - c(t)$, where $c'(t) = -\log d/\alpha$.

As $P'(s) = c'(s)$ for any $s \in \mathbf{R}$, we remark that t is the same one obtained in part (a) of this proof.

Therefore, $I(\beta) = -t\int \log|f'(z)| d(u(t))(z) - h(u(t)) + t\int \log|f'(z)| d(u(t))(z) + \log d = \log d - h(u(t))$.

Therefore

$$\lim_{n\to\infty} \log P_n\{n^{-1} W_n \in (-\log d\alpha^{-1} - \xi, -\log d\alpha^{-1} + \xi)\}$$

is approximately equal to $h(u(t)) - \log d$. In this case

$$d^{-n} \# \{z_i^n | i \in \{1, \cdots, d^n\}, n^{-1} \log |f^{n'}(z_i^n)|^{-1} \in (-\log d\alpha^{-1} - \xi, -\log d\alpha^{-1} + \xi)\}$$

is of order $\exp((h(u(t)) - \log d)n)$, and finally

$$\# \left\{ z_i^n | i \in (1, \cdots, d^n\} \text{ and } |f_i^{n'}(z_i^n)|^{-1} \in \left(\exp\left(\left(-\frac{\log d}{\alpha} - \xi\right)n\right), \exp\left(\left(-\frac{\log d}{\alpha} + \xi\right)n\right)\right)\right\}$$

is of order

$$(*) \qquad \exp(h(u(t)) - (\log d)n) \exp(\log d^n) = \exp(h(u(t))n).$$

As mentioned in § 2, this information allows us to control the number of points with a certain deviation of the mean.

Now we will state some properties of hyperbolic rational maps that are proved in [8] and [18].

Considering perhaps a finite iterate of f, we know from [18] that there exists a curve δ containing all the critical values of f such that:

(a) $u(\delta) = 0$.

(b) $\hat{X} = \mathbf{C} - \delta$ is a topological disk.

(c) There exist branches $\phi_i : \hat{X} \to \bar{\mathbf{C}}$, $i = 1, \cdots, d$, of $f^{-1}|\hat{X}$ that are injective and $f(\hat{X}_i) = \hat{X}$ where $\hat{X}_i = \phi_i(\hat{X})$.

(d) The set $X = (\cap_{n\geqq 0} f^{-n}(\cap_{m\geqq 0} f^m(\hat{X}^c)))^c$ has u-measure zero, and satisfies

$$f^{-1}(X) = X.$$

(e) Set $X_i = \hat{X}_i \cap X$; then the disjoint union $X = \cup_{i=1}^d X_i$ is such that if $n \leqq 1$, $1 \leqq i_j \leqq d$, $j = 1, \cdots, n$, we have d^n sets of the form $(\cap_{j=1}^n f^{-j}(X_{i_j}))$. Let us denote each such set by A_i^n, $i \in \{1, \cdots, d^n\}$.

From [18] we have $u(A_i^n) = d^n$.

(f) We can suppose there exists just one z_i^n in each A_i^n because we can obtain u as $\lim_{n\to\infty} u(n, z_0)$ and this limit does not depend on z_0.

Now let us return to the proof of the theorem. First we will show that $J(\alpha) \cap X$ has dimension smaller than $\mathrm{HD}(u(t))$, where X depends on the curve δ. Then we move the curve δ a little and we obtain the same result. By the injectivity (c) we have that these $J(\alpha) \cap X$ cover $J(\alpha)$ when we consider several different disjoint curves δ, and from this it follows that $f(\alpha) \leqq \mathrm{HD}(u(t))$.

Now we will show that $J(\alpha) \cap X$ has dimension smaller than $\mathrm{HD}(u(t))$ for any curve δ. This will be obtained in the following way. Consider a conformal representation $\phi : X \to D_1$ and $X(r) = \phi^{-1}(D_r)$ (where $D_r = \{z \in \mathbf{C} \mid |z| < r\}$, $0 \leqq r \leqq 1$). Consider also for each A_i^n, $i \in \{1, 2, \cdots, d^n\}$, the corresponding $A_i^n(r)$ such that $A_i^n(r) = f^{-n}(X_r)$ for some branch f^{-n} and $A_i^n(r) \subset A_i^n$, and assume $\phi(z_i^n) = 0$.

We will first show $J(\alpha) \cap \{\cap_{n\geqq 0} f^{-n}(\hat{X}(r))\}$ has dimension smaller than $HD(u(t))$. Note that $J(\alpha)$ is invariant by f. In this case, using the same proof presented in [32] for Theorem 4 and in [25] for Theorem 1.1, we conclude that $\lim_{r\to 1} \mathrm{HD}(J(\alpha) \cap (\cap_{n\geqq 0} f^{-n}(\hat{X}(r)))) = \mathrm{HD}(J(\alpha) \cap X) \leqq \mathrm{HD}(u(t))$. Therefore, it is enough to show that $\mathrm{HD}(J(\alpha) \cap (\cap_{n\geqq 0} f^{-n}(\hat{X}(r))) \leqq \mathrm{HD}(u(t))$, and we will show this now.

From the distortion theorem for univalent functions [8], there exist $c_r, C_r > 0$ such that for n large enough

$$(**) \qquad c_r < |(f^n)'(t)|\,|(f^n)'(z)|^{-1} < C_r$$

for any t, z in $A_i^n(r)$. It also follows from [8] that for any $\xi > 0$, there exists $K > 0$ such that for n large enough, if $D(n, i, \xi)$ is the ball of center z_i^n and radius $K|(f^n)'(z_i^n)|^{-(1-\xi)}$, then

$$(***) \qquad D(n, i, \xi) \supset A_i^n.$$

Consider $Y = J(\alpha) \cap X$. Then for each $z \in Y \cap f^{-n}(\hat{x}(r))$ such that $|(f^n)'(z)|^{-\alpha}$ is of order d^{-n} we have that z is in a certain $A_{i_j}^n(r)$ and therefore from (**)

$$|(f^n)'(z_i^n)|^{-\alpha} \text{ and } |(f^n)'(z)|^{-\alpha} \text{ are of order } d^{-n}.$$

The cardinal of such possible z_i^n is of the order $\exp(h(u(t))n)$ from (*). It is also true that such z is in $D(n, i, \xi)$ from (***).

Now let us remember some properties of $\mathrm{HD}(Y)$. From each $\mathbf{T} > 0$ consider

$$\mathrm{HD}_{\mathbf{T}}(Y) = \lim_{\delta\to 0} \mathrm{HD}_{\mathbf{T},\delta}(Y) = \inf_{\substack{Y \subset \cup B_i \\ \operatorname{diam} B_i \leqq \delta}} \sum (\operatorname{diam} B_i)^T$$

where B_i are balls in $\mathbf{C}$.

We also know that if for all $\mathbf{T} > \mathrm{HD}(u(t))$ we have $H_{\mathbf{T}}(Y)$ finite, then $\mathrm{HD}(Y) \leqq \mathrm{HD}(u(t))$.

Now observe that for each $n \in \mathbf{N}$, Y is contained in $\exp(h(u(t)n))$ balls of radius $K|(f^n)'(z_i^n)|^{-(1-\xi)}$, and $|(f^n)'(z_i^n)|^{-1}$ is of order $\exp(-\log d \cdot n \cdot \alpha^{-1})$. For each n the sum

$$\sum (\operatorname{diam} D(n, i, \varepsilon))^T \leqq K \exp(h(u(t))n) \exp(-T \log d \cdot n \cdot \alpha^{-1}(1-\varepsilon))$$
$$= K \exp(h(u(t)) - \mathbf{T} \log d\alpha^{-1}(1-\varepsilon)n).$$

For

$$\mathbf{T} > \mathrm{HD}(u(t)) = h(u(t)) \left(\int \log |f'(t)| du(t) \right)^{-1} = -h(u(t)) P'(t)^{-1}$$

$$= h(u(t)) \cdot \alpha (\log d)^{-1}$$

we have that the above sum is uniformly bounded. As the diameter $D(n, i, \varepsilon)$ goes to zero [8] because f is hyperbolic, we have

$$\mathrm{HD}\left(J(\alpha) \cap \left(\bigcap_{n \geqq 0} f^{-n}(\hat{X}(r)) \right) \right) \leqq \mathrm{HD}(u(t))$$

and finally the theorem is proved.

The analyticity of $f(\alpha)$ follows from the analyticity of $P(t)$ [23], [28].

As we have mentioned in the Introduction, no dimensionality is lost by considering the support of $u(t)$ instead of $J(\alpha)$ because, as we have just shown, the two sets have the same Hausdorff dimension.

Acknowledgments. We thank P. Collet for supplying us with references on the subject and J. Yorke for some helpful conversations on the topic.

REFERENCES

[1] M. Barnsley and A. N. Harrington, *Moments of balanced measures on Julia sets*, Trans. Amer. Math. Soc., 284 (1984), pp. 271-280.

[2] T. Bohr and D. Rand, *The entropy function for characteristic exponents*, Phys. D, to appear.

[3] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphism*, Lecture Notes in Mathematics 470, Springer-Verlag, Berlin, New York, 1975.

[4] H. Brolin, *Invariant sets under iteration of rational functions*, Ark. Mat. (Band G), 6 (1966), pp. 103-144.

[5] P. Collet, J. Lebowitz, and A. Porzio, *The dimension spectrum of some dynamical systems*, J. Statist. Phys., 47 (1984), pp. 609-644.

[6] A. Douday and J. Hubbard, *Iteration des polynomes quadratiques complexes*, C. R. Acad. Sci. Paris, 294 (1982), pp. 123-126.

[7] R. S. Ellis, *Entropy, Large Deviation and Statistical Mechanics*, Springer-Verlag, Berlin, New York, 1985.

[8] A. Freire, A. Lopes, and R. Mañé, *An invariant measure for rational maps*, Bol. Soc. Brasil Mat., 14 (1983), pp. 45-62.

[9] C. Grebogi, E. Ott, and J. Yorke, *Unstable periodic orbits and the dimension of chaotic attractors*, Phys. Rev. A (3), 7 (1988), pp. 1711-1724.

[10] T. Halsey, M. Jensen, L. Kadanoff, I. Procaccia, and B. Shraiman, *Fractal measures and their singularities: the characterization of strange sets*, Phys. Rev. A (3), 33 (1986), pp. 1141-1151.

[11] H. G. Hentschel and I. Procaccia, *The infinite number of generalized dimensions of fractals and strange attractors*, Phys. D, 8 (1983), pp. 435-444.

[12] M. Jensen, L. Kadanoff, A. Libchaber, I. Procaccia, and J. Stavans, *Global universality at the onset of chaos: results of a forced Rayleigh-Bénard experiment*, Phys. Rev. Lett., 55 (1985), pp. 2798-2810.

[13] A. Lopes, *Equilibrium measures for rational maps*, Ergodic Theory Dynamical Systems, 6 (1986), pp. 393-399.

[14] ———, *Orthogonality and the Hausdorff dimension of the maximal measure*, Proc. Amer. Math. Soc., 98 (1986), pp. 51-55.

[15] ———, *The complex potential generated by the maximal measure for a family of rational maps*, J. Statist. Phys., 52 (1988), pp. 571-575.

[16] ———, *Entropy and large deviation*, preprint.

[17] V. LUBITSH, *Entropy properties of rational endomorphisms on the Riemann sphere*, Ergodic Theory Dynamical Systems, 3 (1983), pp. 351-383.

[18] R. MAÑÉ, *On the Bernoulli property for rational maps*, Ergodic Theory Dynamical Systems, 5 (1985), pp. 71-88.

[19] ———, *Ergodic Theory and Differentiable Dynamics*, Springer-Verlag, Berlin, New York, 1987.

[20] ———, *On the Hausdorff Dimension of the Invariant Probabilities of Rational Maps*, Lecture Notes in Mathematics 1311, Springer-Verlag, Berlin, New York, 1988.

[21] R. MAÑÉ, P. SAD, AND D. SULLIVAN, *On the dynamics of rational maps*, Ann. Sci. École Norm. Sup. (4), 16 (1982), pp. 193-217.

[22] A. MANNING, *The dimension of the maximal measure for polynomial maps*, Ann. of Math. (2), 119 (1984), pp. 425-430.

[23] ———, *A relation between Liapunov exponents, Hausdorff dimension and entropy*, Ergodic Theory Dynamical Systems, 1 (1981), pp. 451-460.

[24] H. MCCLUSKEY AND A. MANNING, *Hausdorff dimension for horseshoes*, Ergodic Theory Dynamical Systems, 3 (1983), pp. 251-260.

[25] L. MENDOZA, *Continuity properties of invariant sets of one-dimensional maps*, Escuela de Ciencias, Cadis, Ucola-Barquisimeto, Venezuela, preprint.

[26] E. OTT, W. WHITTERS, AND J. A. YORKE, *Is the dimension of chaotic attractors invariant under change of coordinates?*, J. Statist. Phys., 36 (1984), pp. 687-703.

[27] D. RAND, *The singularity spectrum for hyperbolic cantor sets and attractors*, Warwick University, preprint.

[28] D. RUELLE, *Repellers for real analytic maps*, Ergodic Theory Dynamical Systems, 2 (1982), pp. 99-107.

[29] ———, *Thermodynamic Formalism*, Addison-Wesley, Reading, MA, 1978.

[30] ———, *Statistical Mechanics*, W.A. Benjamin, New York, 1969.

[31] M. TSUJI, *Potential Theorem in Modern Function Theory*, Maruzen, Tokyo, 1975.

[32] M. URBANSKI, *Invariant subsets of expanding mappings of the circle*, Ergodic Theory Dynamical Systems, 7 (1987), pp. 627-645.

[33] P. WALTERS, *An Introduction to Ergodic Theory*, Springer-Verlag, Berlin, New York, 1982.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1255-1259, September 1989

014

SUPERADDITIVE FUNCTIONS AND A STATISTICAL APPLICATION*

S. Y. TRIMBLE†, JIM WELLS‡, AND F. T. WRIGHT§

Abstract. It is shown that the reciprocals of a class of Laplace transforms are superadditive. This is used to establish the uniqueness of maximum likelihood estimates for a statistical inference problem.

Key words. superadditive, Laplace transform, gamma function, maximum likelihood estimates

AMS(MOS) subject classifications. 26D15, 33A15, 62H12

1. Introduction. In this paper we show that, for $p \geqq 2$ and for x and y in $(0, \infty)$,

$$\frac{1}{\sum_{k=0}^{\infty}(1/(x+k)^p)} + \frac{1}{\sum_{k=0}^{\infty}(1/(y+k)^p)} < \frac{1}{\sum_{k=0}^{\infty}(1/(x+y+k)^p)}. \tag{1}$$

This will follow as a special case of the superadditivity of a class of reciprocals of Laplace transforms, as established in Theorem 2. The special case of (1) when $p=2$ is used to establish the strict concavity of $\log[L(r, s)]$, where

$$L(r, s) := \left[\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} G_1^{r-1} G_2^{s-1}\right]^n.$$

(Here, Γ denotes the gamma function, G_1 and G_2 are both positive, $G_1+G_2<1$, and n is an integer.) This, in turn, will establish the uniqueness of the maximum likelihood estimates of the parameters in the beta distribution.

First, we present a new and simple characterization of a subclass of the completely monotonic functions.

2. Strongly completely monotonic functions. Recall that a real-valued function g is *completely monotonic on* $(0, \infty)$ if and only if $(-1)^n g^{(n)}(x) \geqq 0$, where $0<x<\infty$ and $n=0, 1, 2, \cdots$. We are interested in the more restrictive condition:

$$\begin{aligned}&(-1)^n x^{n+1} g^{(n)}(x) \text{ is nonnegative and nonincreasing for } 0<x<\infty\\ &\text{and } n=0, 1, \cdots.\end{aligned} \tag{2}$$

Such functions will be called *strongly completely monotonic on* $(0, \infty)$. As an example, the function $g(x)=1/x^2$ is strongly completely monotonic in $(0, \infty)$; however, the function $g(x)=e^{-x}$ is completely monotonic, but not strongly completely monotonic.

THEOREM 1. *The function g is strongly completely monotonic on* $(0, \infty)$ *if and only if*

$$g(x) = \int_0^{\infty} e^{-xt}\phi(t)\, dt, \tag{3}$$

where ϕ *is nonnegative and nondecreasing, and where the integral converges for all x in* $(0, \infty)$.

* Received by the editors September 8, 1987; accepted for publication (in revised form) October 31, 1988.

† Department of Mathematics and Statistics, University of Missouri, Rolla, Missouri 65401. Part of this research was done while this author was on leave at the University of Kentucky, Lexington, Kentucky.

‡ Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506. This research was performed during the author's sabbatical at Texas Tech University, Lubbock, Texas 79409.

§ Department of Mathematics and Statistics, University of Missouri, Rolla, Missouri 65401. This author's research was partly supported by Office of Naval Research contract N00014-80-C0322.

Proof. If g is given by (3), then

$$(4)\qquad \begin{aligned}(-1)^n x^{n+1} g^{(n)}(x) &= x^{n+1}\int_0^\infty t^n e^{-xt}\phi(t)\,dt \\ &= \int_0^\infty \tau^n e^{-\tau}\phi\left(\frac{\tau}{x}\right)d\tau \geqq 0,\end{aligned}$$

because $\phi(x)\geqq 0$. Furthermore, the left side is a nonincreasing function of x because $\phi(x)$ is nondecreasing.

Conversely, suppose g is strongly completely monotonic on $(0,\infty)$. Then g is completely monotonic and, according to a representation theorem of Bernstein [5, p. 161], there is a nondecreasing (not necessarily bounded) α such that

$$g(x)=\int_0^\infty e^{-xt}\,d\alpha(t) \qquad (0<x<\infty).$$

Furthermore, we may assume that α is normalized, that is, $\alpha(0)=0$ and $\alpha(x)=(\alpha(x-)+\alpha(x+))/2$ for $0<x<\infty$ [5, pp. 13-14]. Using the Post-Widder inversion formula [5, p. 290], it follows that

$$\alpha(t)-\alpha(0+)=\lim_{n\to\infty}\int_0^t (-1)^n\left(\frac{n}{u}\right)^{n+1} g^{(n)}\left(\frac{n}{u}\right)du$$

for all t in $(0,\infty)$. From (2), it follows that the integrands are nondecreasing functions of u. Hence the integrals, and consequently α itself, are convex functions of t. The properties of convex functions imply that there is a nonnegative and nondecreasing function ϕ such that $\alpha'(t)=\phi(t)$ except, possibly, for a countable set on $(0,\infty)$. Representation (3) follows. □

3. A class of superadditive functions. Let

$$\mathcal{F}=\{f: f \text{ is a continuous, real-valued, nonnegative function defined on } [0,\infty) \text{ satisfying } f(0)=0\}.$$

If $f\in\mathcal{F}$, then f is said to be *superadditive* provided that, for x and y in $(0,\infty)$,

$$(5)\qquad f(x+y)\geqq f(x)+f(y).$$

Beckenback [1, p. 424] remarked that "tests for superadditivity appear to be difficult to establish, and more difficult to apply." In this section and the next, we establish a test for superadditivity and apply it to a problem in statistical inference.

If $f\in\mathcal{F}$, then f is said to be *star-shaped* provided that, for α in $(0,1)$ and x in $(0,\infty)$,

$$f(\alpha x)\leqq \alpha f(x).$$

This is equivalent to saying that

$$(6)\qquad f(x)/x \text{ is nondecreasing on } (0,\infty).$$

It follows from (6) that

$$f(x+y)=x\frac{f(x+y)}{x+y}+y\frac{f(x+y)}{x+y}\geqq x\frac{f(x)}{x}+y\frac{f(y)}{y}=f(x)+f(y).$$

Hence, if f is star-shaped, it is also superadditive. (See Hardy, Littlewood, and Polya [4, p. 83] and Bruckner and Ostrow [2, pp. 1207-1209].) We remark that, if $f(x)/x$ is strictly increasing on $(0,\infty)$, then strict inequality always holds in (5).

THEOREM 2. *Suppose g is strongly completely monotonic and $xg(x)$ does not reduce to a constant, i.e., ϕ is not constant in* (3). *Define f by $f(0)=0$ and $f(x)=1/g(x)$ if $0<x<\infty$. Then f is star-shaped and therefore superadditive. Furthermore, strict inequality always holds in* (5).

Proof. Let $n=0$ in (2). Then $xg(x)$ is nonnegative and nonincreasing, and (6) holds for f. An analysis of (4) shows that $xg(x)$ is strictly decreasing if and only if ϕ is not a constant function. The remarks after (6) complete the proof of this theorem. □

To illustrate, suppose $p \geqq 2$. For $t>0$, we define

$$\phi_p(t) = t^{p-1}/(1-e^{-t}).$$

If we set $\phi_p(0)=0$ for $p>2$ and $\phi_2(0)=1$, it follows that ϕ_p is a continuous, increasing function on $[0,\infty)$. Furthermore, if $x>0$, the Lebesgue monotone convergence theorem implies that

$$\begin{aligned}\int_0^\infty e^{-xt}\phi_p(t)\,dt &= \int_0^\infty \sum_{k=0}^{\infty} t^{p-1}\,e^{-(x+k)t}\,dt \\ &= \sum_{k=0}^{\infty}\int_0^\infty t^{p-1}\,e^{-(x+k)t}\,dt \\ &= \Gamma(p)\sum_{k=0}^{\infty}\frac{1}{(x+k)^p}.\end{aligned}$$

An application of Theorem 2 yields (1) in the Introduction.

It is worth observing that (1) fails for p in the interval $(1,2)$. Indeed, if $1<p<2$, it can be shown that there is a positive number $n(p)$ such that (1) is reversed if x and y are in the interval $(n(p),\infty)$.

In the next section, we shall need the special case of (1) when $p=2$. We then have an inequality concerning the psi (digamma) function, $\psi(x)=\Gamma'(x)/\Gamma(x)$, since $\psi'(x)=\sum_{k=0}^{\infty}1/(x+k)^2$. Actually, $1/\psi'(x)$ is more than superadditive: it is convex on $(0,\infty)$. This fact, which we shall not need in the sequel, follows from a rather tedious argument that establishes the inequality $(\psi'')^2 \geqq \frac{1}{2}\psi'\psi'''$. This derivative inequality is equivalent to

$$\sum_{k=0}^{\infty}\frac{1}{(x+k)^3} \geqq \frac{\sqrt{3}}{2}\left(\sum_{k=0}^{\infty}\frac{1}{(x+k)^2}\right)^{1/2}\left(\sum_{k=0}^{\infty}\frac{1}{(x+k)^4}\right)^{1/2},$$

an inequality of the form

$$\sum a_k b_k \geqq C\sqrt{\sum a_k^2}\,\sqrt{\sum b_k^2}$$

with $a_k=1/(x+k)$, $b_k=1/(x+k)^2$, and $C=\sqrt{3}/2$. Thus the convexity of $1/\psi'(x)$ may be formulated as an inverse Cauchy-Schwartz inequality.

4. A statistical inference problem. Suppose we model a random phenomenon using a beta distribution that has probability density function given by

$$f(x;r,s)=\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}x^{r-1}(1-x)^{s-1}, \qquad 0 \leqq x \leqq 1.$$

Here, r and s are positive numbers to be chosen later. Suppose that $X_1, X_2, \cdots, X_n$ is a random sample. The likelihood for this sample is

$$L(r,s;x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n} f(x_i;r,s)=\left[\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}G_1^{r-1}G_2^{s-1}\right]^n, \tag{7}$$

where $G_1=[\prod_{i=1}^n x_i]^{1/n}$ and $G_2=[\prod_{i=1}^n (1-x_i)]^{1/n}$.

One way of choosing r and s is to find their values that maximize the right side of (7), these being the maximum likelihood estimates. If $G_1 = 0$ or $G_2 = 0$, the right side of (7) can be made infinite by choosing r or s less than 1. So it is natural to assume that $G_1 > 0$ and $G_2 > 0$, which is not a severe restriction because the probability of this happening is 1.

Gnanadesikan, Pinkham, and Hughes [3] have studied the maximum likelihood estimates for (7) but have not addressed the question of the existence and uniqueness of such a solution. This question is important because, in practice, the maximum likelihood estimates are often found by using iterative techniques. Gnanadesikan, Pinkham, and Hughes do observe that $G_1 + G_2 \leqq 1$ with equality if and only if $x_1 = x_2 = \cdots = x_n$, which occurs with probability 0. So, we could assume

$$G_1 + G_2 < 1, \tag{8}$$

which, like our earlier restrictions on G_1 and G_2, would not be severe. In fact, the next result shows that there is also a mathematical reason for assuming (8).

THEOREM 3. *The logarithm of the likelihood function* (7) *is strictly concave on* $(0, \infty) \times (0, \infty)$. *Hence, it has at most one local extremum, which must, if it exists, be an absolute maximum. In fact, the likelihood function has an absolute maximum if and only if* (8) *holds.*

Proof. We write the left side of (7) simply as $L(r, s)$, and view it as defined on $(0, \infty) \times (0, \infty)$. The Hessian matrix of $-\log L(r, s)^{1/n}$ is

$$\begin{bmatrix} \psi'(r) - \psi'(r+s) & -\psi'(r+s) \\ -\psi'(r+s) & \psi'(s) - \psi'(r+s) \end{bmatrix}.$$

The diagonal elements of this matrix are positive on $(0, \infty) \times (0, \infty)$ because ψ' is strictly decreasing, and the determinant is positive on $(0, \infty) \times (0, \infty)$ because of (1). Hence, the matrix is positive definite. Consequently, $\log L(r, s)$ is strictly concave on $(0, \infty) \times (0, \infty)$ and has a local extremum at no more than one point. If this extremum exists, it is necessarily an absolute maximum.

A point will produce a local extremum of $L(r, s)$ if and only if the partial derivatives of $\log L(r, s)$ are both zero there. This requirement yields the following system:

$$\begin{aligned} \psi(r) - \psi(r+s) &= \log G_1, \\ \psi(s) - \psi(r+s) &= \log G_2. \end{aligned} \tag{9}$$

To investigate (9), we establish (10) and (11) below.

Suppose that g is a function defined on $(0, \infty)$ such that $r \leqq g(r)$ for all r. Then $g(r) = r + n(r) + \theta(r)$, where $n(r)$ is a nonnegative integer and $0 \leqq \theta(r) < 1$. Now $\psi(x+1) = \psi(x) + (1/x)$ for $x > 0$. So, for q a nonnegative integer, $\psi(x+q) = \psi(x) + \sum_{k=0}^{q-1} 1/(x+k)$. (If $q = 0$, define $\sum_{k=0}^{q-1} 1/(x+k) = 0$.) If $r \geqq 1$, it follows that

$$\begin{aligned} \log \frac{g(r)-1}{r} &< \log \frac{g(r)-\theta(r)}{r} = \log \frac{r+n(r)}{r} \\ &= \int_0^{n(r)} \frac{dx}{r+x} \leqq \sum_{k=0}^{n(r)-1} \frac{1}{r+k} = \psi(r+n(r)) - \psi(r) \\ &\leqq \psi(g(r)) - \psi(r). \end{aligned} \tag{10}$$

In a similar manner, if $r > 1$,

$$\psi(g(r)) - \psi(r) < \log \frac{g(r)}{r-1}. \tag{11}$$

Without loss of generality, suppose $\log G_2 \geqq \log G_1$. Subtracting the equations in (9) we conclude that, if r and s satisfy (9), we must have

$$\psi(s) - \psi(r) = \log G_2/G_1 \geqq 0. \tag{12}$$

Since ψ is strictly increasing, (12) can be regarded as defining a function $s(r)$ that satisfies $s(r) \geqq r$ for r in $(0, \infty)$. Then (10)-(12) imply that

$$\lim_{r\to\infty} \frac{s(r)}{r} = \frac{G_2}{G_1}. \tag{13}$$

Further, (10), (11), and (13) imply that

$$\lim_{r\to\infty} [\psi(r+s(r)) - \psi(r)] = \log(1 + G_2/G_1). \tag{14}$$

Now $\psi'(x+1) = \psi'(x) - 1/x^2$. Since $\psi'(x+1) > 0$, it follows that $\psi'(x) > 1/x^2$. So $\psi(r+s(r)) - \psi(r) \geqq \psi(2r) - \psi(r) = \int_r^{2r} \psi'(x)\,dx > 1/(2r)$. Hence,

$$\lim_{r\to 0+} [\psi(r+s(r)) - \psi(r)] = +\infty. \tag{15}$$

From (14) and (15), we conclude that the equation

$$\psi(r+s(r)) - \psi(r) = -\log G_1 \tag{16}$$

is satisfied for some r if $-\log G_1 > \log(1 + G_2/G_1)$, i.e., $1 > G_1 + G_2$. Furthermore, if $\psi(r+s(r)) - \psi(r)$ is a strictly decreasing function of r on $(0, \infty)$, then (16) is satisfied for some r if and only if $1 > G_1 + G_2$, i.e., if and only if (8) holds. Since (12) and (16) are equivalent to (9), it only remains to show that $\psi(r+s(r)) - \psi(r)$ is strictly decreasing. But using (1), (12), and the implicit function theorem, we see that

$$\frac{d}{dr}[\psi(r+s(r)) - \psi(r)] = \psi'(r)\psi'(r+s(r))\left[\frac{1}{\psi'(r)} + \frac{1}{\psi'(s(r))} - \frac{1}{\psi'(r+s(r))}\right] < 0. \qquad \square$$

REFERENCES

[1] E. F. BECKENBACK, *Superadditivity inequalities*, Pacific J. Math., 14 (1964), pp. 421-438.

[2] A. M. BRUCKNER AND E. OSTROW, *Some function classes related to the class of convex functions*, Pacific J. Math., 12 (1962), pp. 1203-1215.

[3] R. GNANADESIKAN, R. S. PINKHAM, AND L. P. HUGHES, *Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics*, Technometrics, 9 (1967), pp. 607-620.

[4] G. H. HARDY, J. E. LITTLEWOOD, AND G. POLYA, *Inequalities*, Second edition, Cambridge University Press, London, 1952.

[5] D. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1941.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1260-1269, September 1989

015

APPLICATIONS OF A TECHNIQUE FOR EVALUATING INDEFINITE INTEGRALS CONTAINING PRODUCTS OF THE SPECIAL FUNCTIONS OF PHYSICS*

JEAN C. PIQUETTE†

Abstract. An earlier article [J. C. Piquette and A. L. Van Buren, *SIAM J. Math. Anal.*, 15 (1984), pp. 845-855] described an analytical technique for evaluating indefinite integrals involving special functions. The technique replaces the integral by an inhomogeneous set of coupled first-order differential equations. This coupled set does not explicitly contain the special functions of the integrand, and any particular solution of the set is sufficient to obtain an analytical result for the indefinite integral. One difficulty with the technique is that the process of uncoupling one of the functions of the set from the remainder is not straightforward and, in some instances, may be intractable. In the present article, the relevant set is uncoupled for a number of special cases in which the integrand contains one or more Bessel functions, Legendre functions, Hermite functions, or Laguerre functions. The utility of the given analytical expressions is demonstrated by presenting several examples. Several of the indefinite integrals evaluated have not been previously tabulated.

Key words. integration, antidifferentiation, special functions, systems of differential equations

AMS(MOS) subject classifications. 26A36, 33A40, 33A45, 33A65

1. Introduction. A previous article [1] presented an analytical technique for evaluating indefinite integrals of the form

$$I = \int dx f(x) \prod_{i=1}^{m} R_{\mu_i}^{(i)}(x) \tag{1}$$

where $R_{\mu_i}^{(i)}(x)$ is of the ith type of special functions of order μ_i obeying the set of recurrence relations

$$R_{\mu+1}^{(i)}(x) = a_\mu(x) R_\mu^{(i)}(x) + b_\mu(x) R_{\mu-1}^{(i)}(x), \tag{2a}$$

$$DR_\mu^{(i)}(x) = c_\mu(x) R_\mu^{(i)}(x) + d_\mu(x) R_{\mu-1}^{(i)}(x). \tag{2b}$$

Here a_μ, b_μ, c_μ, and d_μ are known functions corresponding to $R_\mu^{(i)}$. The symbol D represents d/dx. The function $f(x)$ and the product $\prod R_\mu^{(i)}$ both are assumed continuous (or with at most a finite number of discontinuities) over an interval $[x_1, x_2]$, ensuring that the integral I exists in the same interval. The technique is a generalization of one used by Sonine [2] (described by Watson [3]) to evaluate indefinite integrals involving products of Bessel functions. Piquette and Van Buren [1] extended the technique to include most of the special functions of physics, including Legendre functions, Hermite functions, and Laguerre functions. The technique replaces the integral that is to be evaluated by an inhomogeneous set of coupled first-order differential equations. The coupled set does not explicitly contain any of the special functions $R_\mu^{(i)}$, and *any particular solution* of the set is sufficient to yield an analytical expression for the integral I. Here we use the term "birecurrent functions" that was introduced in [1] for the functions $R_\mu^{(i)}$.

The method of [1] assumes that the integral I of (1) may be represented by

$$I = \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \cdots \sum_{p_m=0}^{1} A_{p_1 p_2 \cdots p_m}(x) \prod_{i=1}^{m} R_{(\mu_i+p_i)}^{(i)}(x), \tag{3}$$

* Received by the editors March 9, 1987; accepted for publication (in revised form) November 18, 1988.

† Naval Research Laboratory, Underwater Sound Reference Detachment, P.O. Box 568337, Orlando, Florida 32856-8337.

where the 2^m coefficients $A_{p_1 p_2 \cdots p_m}(x)$ are functions to be determined. The technique replaces the integral I with the coupled set of differential equations

$$f(x)\delta_{0,p} = DA_p + \sum_{\{q\}} B_{pq} A_q, \tag{4}$$

where δ is a Kronecker delta defined to be zero unless $p_1 = p_2 = \cdots = p_m = 0$. In (4), the shorthand notation $A_p = A_{p_1 p_2 \cdots p_m}(x)$ and $B_{pq} = B_{p_1, p_2, \cdots, p_m, q_1, q_2, \cdots, q_m}(x)$ has been used. Also, the notation $\sum_{\{q\}}$ represents the multiple summations $\sum_{q_1=0}^{1} \sum_{q_2=0}^{1} \cdots \sum_{q_m=0}^{1}$. The functions B_{pq} are known functions resulting from repeated applications of the relations of (2) and the regrouping of terms in the form $\prod_{i=1}^{m} R_{\mu_i + p_i}^{(i)}$. An analytical expression for the functions B_{pq} appears in a subsequent publication [4].

One difficulty with the technique is that the process of uncoupling one function A_p from the remainder of the set is not straightforward, and may be intractable in certain cases. To increase the general utility of the method, the present article provides uncoupled equations for several special cases in which the integrand involves one or more Bessel functions, Legendre functions, Hermite functions, or Laguerre functions. The usefulness of the expressions is demonstrated by obtaining solutions to the uncoupled equations in several particular cases, thereby obtaining the associated indefinite integrals. Several of these have not been previously tabulated.

2. Integrands containing one birecurrent function. First we will consider integrals of the general form

$$I = \int dx f(x) R_\mu(x). \tag{5}$$

In this case, the assumed form generated by (3) reduces to

$$I = A_0(x) R_\mu(x) + A_1(x) R_{\mu+1}(x). \tag{6}$$

The resulting coupled set obtained from (4) is

$$\begin{aligned} &-A_0(x)[a_\mu(x)d_\mu(x) - b_\mu(x)c_\mu(x)] \\ &\quad + b_\mu(x)DA_0(x) + b_\mu(x)d_{\mu+1}(x)A_1(x) = b_\mu(x)f(x), \end{aligned} \tag{7a}$$

$$d_\mu(x)A_0(x) + b_\mu(x)DA_1(x) + b_\mu(x)c_{\mu+1}(x)A_1(x) = 0. \tag{7b}$$

Note in this case that, since (7b) does not contain DA_0, this equation can be solved algebraically for A_0 in terms of A_1 and DA_1. This expression may be substituted into (7a) to yield an uncoupled equation for A_1. Since we have used generic expressions for the functions a, b, c, and d, it is clear that the coupled set (7a, b) can always be uncoupled regardless of which particular function R_μ is contained in (5). Therefore, the method always results in a set that can be uncoupled when it is applied to integrals containing a single birecurrent function. Of course, obtaining a particular solution to the resulting uncoupled inhomogeneous differential equation is not straightforward. However, a number of special cases have been considered successfully in [1]. Also, the uncoupled equations for the special cases where R_μ represents either a Bessel, a Legendre, a Hermite, or a Laguerre function are also given in [1]. Therefore, we will not consider this case further in the present article.

3. Integrands containing two Legendre functions. Unlike the case considered in § 2, the coupled set for the case in which the integrand contains two arbitrary birecurrent

functions cannot be uncoupled for the general case in a straightforward way. In this section, we consider the particular case in which the integral of interest is of the form

$$I = \int dx f(x) P_\mu(x) P_\nu(x), \tag{8}$$

where P can be either the first or the second solution to the Legendre differential equation. For the moment, we restrict our attention to cases in which $\mu \neq \nu$. In this case, the general assumed form generated by (3) yields the expression

$$\begin{aligned} I = {} & A_{00}(x) P_\mu(x) P_\nu(x) + A_{01}(x) P_\mu(x) P_{\nu+1}(x) \\ & + A_{10}(x) P_{\mu+1}(x) P_\nu(x) + A_{11}(x) P_{\mu+1}(x) P_{\nu+1}(x). \end{aligned} \tag{9}$$

The function A_{11} can be uncoupled from the coupled set obtained from (4) to give

$$\begin{aligned} & (x^2-1)^2 D^4 A_{11}(x) + 10x(x^2-1) D^3 A_{11}(x) \\ & \quad -2(4-\mu-\nu+\mu x^2+\nu x^2+\mu^2 x^2+\nu^2 x^2-\mu^2-\nu^2-12x^2) D^2 A_{11}(x) \\ & \quad -6x(-2+\mu+\nu+\mu^2+\nu^2) D A_{11}(x) \\ & \quad +(\mu+\nu)(-1+\mu-\nu)(1+\mu-\nu)(2+\mu+\nu) A_{11}(x) \\ & = 2(1+\mu)(1+\nu) D f(x). \end{aligned} \tag{10}$$

The remaining functions A can be expressed in terms of $A_{11}(x)$ and its derivatives. The expressions are

$$\begin{aligned} A_{00}(x) = {} & \frac{-x}{1+\mu+\nu} f(x) - \frac{(x^2-1)}{(\mu-\nu)(1+\mu+\nu)} D f(x) \\ & + g_{000}(x) A_{11}(x) + g_{001}(x) D A_{11}(x) + g_{002}(x) D^2 A_{11}(x) \\ & + g_{003}(x) D^3 A_{11}(x) + g_{004}(x) D^4 A_{11}(x), \end{aligned} \tag{11}$$

$$\begin{aligned} A_{01}(x) = {} & \frac{-(1+\nu)}{(\mu-\nu)(1+\mu+\nu)} f(x) + g_{010}(x) A_{11}(x) + g_{011}(x) D A_{11}(x) \\ & + g_{012}(x) D^2 A_{11}(x) + g_{013}(x) D^3 A_{11}(x), \end{aligned} \tag{12}$$

$$\begin{aligned} A_{10}(x) = {} & \frac{(1+\mu)}{(\mu-\nu)(1+\mu+\nu)} f(x) + g_{100}(x) A_{11}(x) + g_{101}(x) D A_{11}(x) \\ & + g_{102}(x) D^2 A_{11}(x) + g_{103}(x) D^3 A_{11}(x). \end{aligned} \tag{13}$$

The parametric functions g in (11)-(13) are given by

$$\begin{aligned} g_{000}(x) = {} & [\mu+\nu+\mu\nu+\mu\nu^2-2\nu x^2+\mu^2\nu \\ & +\mu^2\nu^2-\nu^2x^2+2\nu^3x^2+\nu^4x^2-\mu\nu x^2-4\mu\nu^2x^2 \\ & -\mu\nu^3x^2+2\mu^2\nu x^2-\mu^2\nu^2x^2+\mu^3\nu x^2+\mu^2-2\nu^3-\nu^4] \\ & \div[(1+\mu)(1+\nu)(\mu-\nu)(1+\mu+\nu)], \end{aligned} \tag{14}$$

$$\begin{aligned} g_{001}(x) = {} & -x[12-12\nu-3\mu\nu^2-4\mu x^2+16\nu x^2 \\ & +3\mu^2\nu+3\mu^2x^2+\mu^3x^2+9\nu^2x^2-\nu^3x^2 \\ & +3\mu\nu^2x^2-3\mu^2\nu x^2-3\mu^2-\mu^3-9\nu^2+\nu^3-12x^2] \\ & \div[2(1+\mu)(1+\nu)(\mu-\nu)(1+\mu+\nu)], \end{aligned} \tag{15}$$

$$(16)\quad g_{002}(x)=-\frac{[(x^2-1)(8-\mu-3\nu-5\mu x^2+9\nu x^2+\mu^2x^2+3\nu^2x^2-\mu^2-3\nu^2-24x^2)]}{2(1+\mu)(1+\nu)(\mu-\nu)(1+\mu+\nu)},$$

$$(17)\quad g_{003}(x)=\frac{x(-1+x^2)^2(10+\mu-\nu)}{2(1+\mu)(1+\nu)(\mu-\nu)(1+\mu+\nu)},$$

$$(18)\quad g_{004}(x)=\frac{(-1+x^2)^3}{2(1+\mu)(1+\nu)(\mu-\nu)(1+\mu+\nu)},$$

$$(19)\quad g_{010}(x)=\frac{-\mu x(1+\mu-\nu)(2+\mu+\nu)}{(1+\mu)(\mu-\nu)(1+\mu+\nu)},$$

$$(20)\quad g_{011}(x)=-\frac{[2-3\mu-\nu+3\mu x^2+\nu x^2+3\mu^2x^2+\nu^2x^2-3\mu^2-\nu^2-6x^2]}{2(1+\mu)(\mu-\nu)(1+\mu+\nu)},$$

$$(21)\quad g_{012}(x)=\frac{3x(-1+x^2)}{(1+\mu)(\mu-\nu)(1+\mu+\nu)},$$

$$(22)\quad g_{013}(x)=\frac{(-1+x^2)^2}{2(1+\mu)(\mu-\nu)(1+\mu+\nu)},$$

$$(23)\quad g_{100}(x)=\frac{\nu x(1-\mu+\nu)(2+\mu+\nu)}{(1+\nu)(\mu-\nu)(1+\mu+\nu)},$$

$$(24)\quad g_{101}(x)=\frac{[2-\mu-3\nu+\mu x^2+3\nu x^2+\mu^2x^2+3\nu^2x^2-\mu^2-3\nu^2-6x^2]}{2(1+\nu)(\mu-\nu)(1+\mu+\nu)},$$

$$(25)\quad g_{102}(x)=\frac{-3x(-1+x^2)}{(1+\nu)(\mu-\nu)(1+\mu+\nu)},$$

$$(26)\quad g_{103}(x)=\frac{-(-1+x^2)^2}{2(1+\nu)(\mu-\nu)(1+\mu+\nu)}.$$

Although the expressions above are complicated, they are straightforward to evaluate, with the exception of (10), the differential equation for A_{11}. However, recall that *any particular solution* of (10) is adequate for obtaining the integral of (8). There are several cases in which this is not difficult to do.

For example, if $f(x)=1$ in (8), the inhomogeneous term in (10) vanishes. Therefore, a particular solution to (10) is simply $A_{11}(x)=0$. Substituting this into (11)-(26) immediately yields the indefinite integral

$$(27)\quad \int dx\, P_\mu(x)P_\nu(x)=\frac{-xP_\mu(x)P_\nu(x)}{1+\mu+\nu}-\frac{(1+\nu)}{(\mu-\nu)(1+\mu+\nu)}P_\mu(x)P_{\nu+1}(x)+\frac{(1+\mu)}{(\mu-\nu)(1+\mu+\nu)}P_{\mu+1}(x)P_\nu(x)$$

when the resulting expressions for A_{00}, A_{01}, and A_{10} are substituted into (9). An integral equivalent to (27) can be found, for example, in [5].

The differential equation (10) can also be solved for most functions $f(x)$ of the form x^l, where l is a natural number, thus yielding the antiderivative

$$(28)\quad I=\int dx\, x^lP_\mu(x)P_\nu(x).$$

A particular solution to (10) for $f(x)=x^l$ can be represented by the truncated series

$$A_{11}(x)=\sum_{p}^{l-1}{}' b_p x^p. \tag{29}$$

In (29), the quantities b_p are constants given by

$$b_{l-1}=\frac{(2l)(1+\mu)(1+\nu)}{(\mu+\nu+l+1)(\mu+\nu-l+1)[(\mu-\nu)^2-l^2]}, \tag{30a}$$

$$b_p=-\frac{(p+1)(p+2)\{(p+3)(p+4)b_{p+4}(1-\delta_{p,l-3})+2b_{p+2}[\mu+\nu+\mu^2+\nu^2-(p+2)^2]\}}{(\mu+\nu+p+2)(\mu+\nu-p)[(\mu-\nu)^2-(p+1)^2]} \tag{30b}$$

$$(0\leqq p<l-1).$$

The prime on the summation in (29) signifies that $p=0, 2, \cdots, l-1$ when l is odd, and $p=1, 3, \cdots, l-1$ when l is even. The particular case $l=0$ is excluded, but note that this case is covered by (27). The orders μ and ν are arbitrary (and may even be complex), except for the particular cases in which the denominators in (30) vanish. The desired antiderivative corresponding to any particular value of l is obtained by substituting the resulting expression for A_{11} into (11)-(26). The results of these calculations are then substituted into (9). This process is difficult to perform for the general case. However, one particular case is

$$\begin{aligned}\int dx\, x^5 P_{1/3}(x)P_{2/3}(x)=&\left(-\frac{335}{2352}-\frac{515x^2}{392}+\frac{235x^4}{336}+\frac{x^6}{3}\right)P_{1/3}(x)P_{2/3}(x)\\&+\left(\frac{685x}{784}-\frac{575x^3}{1176}-\frac{5x^5}{48}\right)P_{1/3}(x)P_{5/3}(x)\\&+\left(\frac{235x}{196}-\frac{295x^3}{588}-\frac{2x^5}{21}\right)P_{2/3}(x)P_{4/3}(x)\\&+\left(-\frac{5}{6}+\frac{65x^2}{196}+\frac{25x^4}{588}\right)P_{4/3}(x)P_{5/3}(x).\end{aligned} \tag{31}$$

The uncoupled equations given by (10)-(26) are not valid if $\mu=\nu$. In this special case a different uncoupling scheme is required. The uncoupled set for this case is given by

$$\begin{aligned}&(-1+x^2)^2D^3A_{11}(x)+6x(-1+x^2)D^2A_{11}(x)\\&\quad-2(1-2\nu+2\nu x^2+2\nu^2x^2-2\nu^2-3x^2)DA_{11}(x)\\&\quad-4\nu(\nu+1)xA_{11}(x)=2(\nu+1)^2f(x),\end{aligned} \tag{32}$$

$$A_{00}(x)=\frac{(-1+x^2)^2}{2(1+\nu)^2}D^2A_{11}(x)+\frac{(-1+x^2)(2+\nu)}{(1+\nu)^2}xDA_{11}(x)+\frac{(\nu+x^2)}{(1+\nu)}A_{11}(x), \tag{33}$$

$$A_{01}(x)=-\frac{(-1+x^2)}{2(1+\nu)}DA_{11}(x)-xA_{11}(x), \tag{34}$$

and $A_{10}(x)=A_{01}(x)$. As one particular example of this case, we let $f(x)=x$; a particular solution of (32) is then $A_{11}=-(1+\nu)/2\nu$. When the resulting solutions obtained from (33) and (34) are substituted into (9), this yields the integral

$$\int dx\, x[P_\nu(x)]^2=\frac{-(1+\nu)}{2\nu}\left[\left(\frac{x^2+\nu}{1+\nu}\right)[P_\nu(x)]^2-2xP_\nu(x)P_{\nu+1}(x)+[P_{\nu+1}(x)]^2\right]. \tag{35}$$

The expressions given in (32)-(34) also apply if the integral of interest is

$$I = \int dx f(x) P_\nu(x) Q_\nu(x), \tag{36}$$

where $Q_\nu(x)$ is the second solution to Legendre's differential equation. This follows from the fact that the functions P and Q each obey the same recurrence relations. As a particular example of this, let $f(x) = x^3$ in (32). A particular solution of this differential equation is then

$$A_{11}(x) = \frac{(1+\nu)(-4+2\nu+\nu x^2+\nu^2x^2+2\nu^2)}{6\nu(1-\nu)(2+\nu)}. \tag{37}$$

Substituting this into (33) and (34), we obtain the integral

$$\begin{aligned}\int dx\, x^3 P_\nu(x) Q_\nu(x) &= -\frac{(-3\nu-4\nu x^2+6\nu x^4+\nu^2x^2+3\nu^2x^4+\nu^3x^2+2\nu^2+2\nu^3-4\nu^2)P_\nu(x)Q_\nu(x)}{6\nu(-1+\nu)(2+\nu)} \\ &\quad + \frac{x(1+\nu)(-4+\nu+2\nu x^2+\nu^2x^2+2\nu^2)}{6\nu(-1+\nu)(2+\nu)}[P_\nu(x)Q_{\nu+1}(x)+P_{\nu+1}(x)Q_\nu(x)] \\ &\quad + \frac{(1+\nu)(-4+2\nu+\nu x^2+\nu^2x^2+2\nu^2)}{6\nu(1-\nu)(2+\nu)} P_{\nu+1}(x)Q_{\nu+1}(x)\end{aligned} \tag{38}$$

when the resulting solutions for the functions A are substituted into the representation of (9).

4. Integrands containing the square of a Bessel function. We now turn our attention to integrals of the form

$$I = \int dx f(x) Z_\nu^2(x), \tag{39}$$

where Z_ν is an arbitrary cylinder function, i.e., any solution of Bessel's differential equation. In this case, (3) provides the representation

$$I = A_{00}(x)Z_\nu^2(x) + 2A_{01}(x)Z_\nu(x)Z_{\nu+1}(x) + A_{11}(x)Z_{\nu+1}^2(x). \tag{40}$$

Once again, we can uncouple A_{11} from the set generated by (4), giving

$$\begin{aligned}&x^3D^3A_{11}(x) - 3x^2D^2A_{11}(x) + x(7-4\nu^2+4x^2)DA_{11}(x) \\ &\quad + 4(-2+2\nu^2-x^2)A_{11}(x) = 2x^3f(x),\end{aligned} \tag{41}$$

$$A_{00}(x) = \frac{1}{2}D^2A_{11}(x) - \frac{(3+2\nu)}{2x}DA_{11}(x) + \frac{(2+2\nu+x^2)}{x^2}A_{11}(x), \tag{42}$$

$$A_{01}(x) = \frac{1}{2}DA_{11}(x) - \frac{(1+\nu)}{x}A_{11}(x). \tag{43}$$

As an example of this class of integrals, we first let $f(x) = 1/x^2$ in (39). Next, we consider the case in which $f(x) = 1/x^4$. The first of these examples results in the integral

$$\int \frac{Z_\nu^2(x)}{x^2}dx = \frac{1+2\nu+2x^2}{(4\nu^2-1)x}Z_\nu^2(x) - \frac{2}{-1+2\nu}Z_\nu(x)Z_{\nu+1}(x) - \frac{2x}{1-4\nu^2}Z_{\nu+1}^2(x), \tag{44}$$

and the second example yields

$$\int \frac{Z_\nu^2(x)}{x^4}\,dx = \frac{[-9-6\nu+x^2(6+16\nu+8\nu^2)+36\nu^2+24\nu^3+16x^4]}{3x^3(1-4\nu^2)(9-4\nu^2)} Z_\nu^2(x)$$

$$-\left[\frac{2(-3+4\nu+4\nu^2+8x^2)}{3x^2(1-2\nu)(9-4\nu^2)}\right] Z_\nu(x)Z_{\nu+1}(x) \tag{45}$$

$$-\frac{2(1-4\nu^2-8x^2)}{3x(1-4\nu^2)(9-4\nu^2)} Z_{\nu+1}^2(x)$$

when the particular solutions to (41) for A_{11}, given by the coefficients of the $Z_{\nu+1}^2(x)$ terms in (44) and (45), are substituted into (42) and (43), and the resulting expressions are used in (40). A result equivalent to (44) can be deduced using [6, p. 257, eq. 21], although that formula is restricted to Bessel functions of the first kind. This integral can also be found in formula (4) of [7, § 1.13.3, p. 50]. To obtain a result equivalent to (45) combine formulas (1) and (4) of [7, § 1.13.3, p. 50].

5. Integrands containing the square of a Hermite function or the square of a Laguerre function. We next consider integrals of the form

$$I = \int dx f(x) H_\nu^2(x), \tag{46}$$

where H_ν is any solution of the Hermite differential equation. In this case, (3) gives the assumed form

$$I = A_{00}(x)H_\nu^2(x) + 2A_{01}(x)H_\nu(x)H_{\nu+1}(x) + A_{11}(x)H_{\nu+1}^2(x). \tag{47}$$

The coupled set resulting from (4) may be uncoupled to yield

$$D^3A_{11}(x) + 6xD^2A_{11}(x) + 2(5+4\nu+4x^2)DA_{11}(x) + 16x(1+\nu)A_{11}(x) = 2f(x), \tag{48}$$

$$A_{00}(x) = \tfrac{1}{2}D^2A_{11}(x) + xDA_{11}(x) + 2(1+\nu)A_{11}(x), \tag{49}$$

$$A_{01}(x) = \tfrac{1}{2}DA_{11}(x). \tag{50}$$

Next consider the integral

$$I = \int dx f(x) L_\nu^2(x), \tag{51}$$

where L_ν is any solution of the Laguerre differential equation. The assumed form of the solution provided by (3) for this case is

$$I = A_{00}(x)L_\nu^2(x) + 2A_{01}(x)L_\nu(x)L_{\nu+1}(x) + A_{11}(x)L_{\nu+1}^2(x). \tag{52}$$

The coupled set generated by (4) may be uncoupled to obtain the equations

$$x^2D^3A_{11}(x) + 3x(1+x)D^2A_{11}(x)$$

$$+(1+8x+4\nu x+2x^2)DA_{11}(x) + 2(1+2x)(1+\nu)A_{11}(x) \tag{53}$$

$$= 2(1+\nu)^2 f(x),$$

$$A_{00}(x) = \frac{x^2}{2(1+\nu)^2} D^2A_{11}(x) + \frac{x(3+2\nu+x)}{2(1+\nu)^2} DA_{11}(x) + \frac{(1+\nu+x)}{1+\nu} A_{11}(x), \tag{54}$$

$$A_{01}(x) = -\frac{x}{2(1+\nu)} DA_{11}(x) - A_{11}(x). \tag{55}$$

6. Integrands containing higher powers of birecurrent functions. In addition to the special cases presented above, the general cases of integrands containing a single *arbitrary* birecurrent function raised to either the second, third, or fourth power have each been uncoupled analytically. It is reasonable to conjecture that integrands of the form

$$I = \int dx f(x)[R_\mu(x)]^n, \tag{56}$$

where n is a natural number, can always be uncoupled. However, an analytical demonstration of this has proved elusive so far.

As one example in this category, we consider the integral

$$I = \int dx f(x) Z_\nu^4(x), \tag{57}$$

where Z_ν is again any solution of Bessel's differential equation. The assumed form generated by (3) in this case yields

$$\begin{aligned} I = {} & A_{0000}(x)Z_\nu^4(x) + A_{1111}(x)Z_{\nu+1}^4(x) + 4A_{0001}(x)Z_\nu^3(x)Z_{\nu+1}(x) \\ & + 6A_{0011}(x)Z_\nu^2(x)Z_{\nu+1}^2(x) + 4A_{0111}(x)Z_\nu(x)Z_{\nu+1}^3(x). \end{aligned} \tag{58}$$

The function $A_{1111}(x)$ can be uncoupled from the set produced by (4) to give

$$\begin{aligned} & x^5 D^5 A_{1111}(x) - 10x^4 D^4 A_{1111}(x) - 5x^3(-13+4\nu^2-4x^2) D^3 A_{1111}(x) \\ & \quad + 15x^2(-19+12\nu^2-8x^2) D^2 A_{1111}(x) \\ & \quad + x(781-128\nu^2x^2-740\nu^2+64\nu^4+392x^2+64x^4) D A_{1111}(x) \\ & \quad - 64(16-6\nu^2x^2-20\nu^2+4\nu^4+9x^2+2x^4) A_{1111}(x) = 24x^5 f(x). \end{aligned} \tag{59}$$

The remaining functions A are expressed in terms of $A_{1111}(x)$ by the equations

$$\begin{aligned} A_{0000}(x) = {} & \frac{1}{24} D^4 A_{1111}(x) - \frac{(5+2\nu)}{12x} D^3 A_{1111}(x) \\ & - \frac{(-55-36\nu+4\nu^2-16x^2)}{24x^2} D^2 A_{1111}(x) \\ & + \frac{(-175-148\nu-40\nu x^2+28\nu^2+16\nu^3-84x^2)}{24x^3} D A_{1111}(x) \\ & + \frac{(32+32\nu+16\nu x^2-8\nu^2-8\nu^3+20x^2+3x^4)}{3x^4} A_{1111}(x), \end{aligned} \tag{60}$$

$$\begin{aligned} A_{0001}(x) = {} & \frac{1}{24} D^3 A_{1111}(x) - \frac{(3+2\nu)}{8x} D^2 A_{1111}(x) \\ & + \frac{(37+42\nu+8\nu^2+10x^2)}{24x^2} D A_{1111}(x) \\ & - \frac{(8+12\nu+3\nu x^2+4\nu^2+4x^2)}{3x^3} A_{1111}(x), \end{aligned} \tag{61}$$

$$A_{0011}(x) = \frac{1}{12} D^2 A_{1111}(x) - \frac{(7+6\nu)}{12x} D A_{1111}(x) + \frac{(4+6\nu+2\nu^2+x^2)}{3x^2} A_{1111}(x), \tag{62}$$

$$A_{0111}(x) = \frac{1}{4} D A_{1111}(x) - \frac{(1+\nu)}{x} A_{1111}(x). \tag{63}$$

In view of the complexity of the differential equation (59), it is difficult to obtain solutions for arbitrary orders ν. Therefore, in this case, we restrict our attention to examples of particular orders.

As one example in this category, we consider $f(x)=1/x$, $\nu=1$. A particular solution to (59) for this case is $A_{1111}(x)=x^2/4$. This produces the integral

$$\begin{aligned}\int dx\frac{Z_1^4(x)}{x}&=\frac{x^2}{4}Z_2^4(x)+\left(\frac{3}{4}+\frac{x^2}{4}\right)Z_1^4(x)-\frac{3x}{2}Z_1(x)Z_2^3(x)\\&\quad+6\left(\frac{1}{2}+\frac{x^2}{12}\right)Z_1^2(x)Z_2^2(x)+4\left(-\frac{3x}{8}-\frac{1}{2x}\right)Z_1^3(x)Z_2(x).\end{aligned}\tag{64}$$

As a second example, we let $f(x)=1/x^3$ and $\nu=3$. A particular solution to (59) can also be obtained for this case, thus yielding the integral

$$\begin{aligned}\int dx\frac{Z_3^4(x)}{x^3}&=\left(\frac{1}{24}+\frac{1}{2x^2}+\frac{2}{x^4}+\frac{x^2}{378}\right)Z_3^4(x)+\left(\frac{5}{216}+\frac{2}{27x^2}+\frac{x^2}{378}\right)Z_4^4(x)\\&\quad+4\left(-\frac{x}{108}-\frac{5}{54x}-\frac{1}{3x^3}\right)Z_3(x)Z_4^3(x)\\&\quad+6\left(\frac{7}{216}+\frac{1}{3x^2}+\frac{4}{3x^4}+\frac{x^2}{1134}\right)Z_3^2(x)Z_4^2(x)\\&\quad+4\left(-\frac{x}{108}-\frac{1}{8x}-\frac{1}{x^3}-\frac{4}{x^5}\right)Z_3^3(x)Z_4(x).\end{aligned}\tag{65}$$

Since the analytical expressions for the uncoupled equations resulting from the special cases when $n=2$, 3, 4 are extremely complicated, they will not be displayed. We will instead provide a tabulation of indefinite integrals obtained by computing particular solutions to these uncoupled equations for certain special cases. In view of the very little previous work that has been done concerning integrals containing more than two special functions, (64), (65), and the following integrals appear to be original tabulations:

$$\begin{aligned}\int dx\,x^2Z_{1/3}^3(x)&=\left(-\frac{4}{9}x-\frac{16}{81x}\right)Z_{1/3}^3(x)-(4x/3)Z_{1/3}(x)Z_{4/3}^2(x)\\&\quad+\left(\frac{8}{9}+x^2\right)Z_{1/3}^2(x)Z_{4/3}(x)+\frac{2}{3}x^2Z_{4/3}^3(x),\end{aligned}\tag{66}$$

$$\begin{aligned}\int dx\,x[P_{1/3}(x)]^3&=\left(\frac{125x^4}{12}-\frac{14}{3}x^2-\frac{5}{12}\right)[P_{1/3}(x)]^3+(-4+20x^2)P_{1/3}(x)[P_{4/3}(x)]^2\\&\quad+(9x-25x^3)[P_{1/3}(x)]^2P_{4/3}(x)-\frac{16}{3}x[P_{4/3}(x)]^3,\end{aligned}\tag{67}$$

$$\begin{aligned}\int dx\,e^{-3x^2}x^2H_{2/3}^3(x)&=e^{-3x^2}\{-\tfrac{1}{12}x(5+6x^2)H_{2/3}^3(x)+\tfrac{1}{8}(1+6x^2)H_{2/3}^2(x)H_{5/3}(x)\\&\quad-\tfrac{3}{8}xH_{2/3}(x)H_{5/3}^2(x)+\tfrac{1}{16}H_{5/3}^3(x)\},\end{aligned}\tag{68}$$

$$\begin{aligned}&\int dx\,e^{-2x^2}x^3H_{-2/3}^3(x)\\&=\frac{e^{-2x^2}}{16}\left\{\frac{9x}{4}H_{1/3}^3(x)+\frac{9}{4}(1-4x^2)H_{1/3}^2(x)H_{-2/3}(x)\right.\\&\quad\left.-9x(1-x^2)H_{1/3}(x)H_{-2/3}^2(x)+\frac{1}{4}(-8+20x^2)H_{-2/3}^3(x)\right\},\end{aligned}\tag{69}$$

$$\int dx\, e^{-3x} x L_{2/3}^3(x)$$

$$= e^{-3x}\left\{ L_{2/3}^3(x)\left[\frac{125}{24}-\frac{625x}{24}+\frac{853x^2}{16}-\frac{675x^3}{16}+\frac{225x^4}{16}-\frac{27x^5}{16}\right]\right.$$

$$+ L_{5/3}^3(x)\left[-\frac{125}{24}+\frac{125x}{12}-\frac{125x^2}{16}\right] \tag{70}$$

$$+3\left[\frac{125}{24}-\frac{125x}{8}+\frac{275x^2}{16}-\frac{75x^3}{16}\right] L_{2/3}(x) L_{5/3}^2(x)$$

$$\left.+3\left[-\frac{125}{24}+\frac{125x}{6}-\frac{515x^2}{16}+\frac{135x^3}{8}-\frac{45x^4}{16}\right] L_{2/3}^2(x) L_{5/3}(x)\right\},$$

$$\int dx\, x[P_{1/2}(x)]^4 = \left(-\frac{5}{16}-\frac{19}{4}x^2\right)[P_{1/2}(x)]^4+\frac{81}{4}xP_{1/2}(x)[P_{3/2}(x)]^3$$

$$+6\left(-\frac{9}{16}-\frac{9}{2}x^2\right)[P_{1/2}(x)]^2[P_{3/2}(x)]^2 \tag{71}$$

$$+4\left(\frac{33x}{16}+3x^3\right)[P_{1/2}(x)]^3 P_{3/2}(x)-\frac{81}{16}[P_{3/2}(x)]^4.$$

In (66)-(71), Z denotes the Bessel function, P the Legendre function, H the Hermite function, and L the Laguerre function.

Note added in proof. Since the time of writing, additional progress has been made on the uncoupling problem. A manuscript [8] describing this work has been submitted for publication.

REFERENCES

[1] J. C. PIQUETTE AND A. L. VAN BUREN, *Technique for evaluating indefinite integrals involving products of certain special functions*, SIAM J. Math. Anal., 15 (1984), pp. 845-855.

[2] N. J. SONINE, *Recherches sur les fonctions cylindriques et le développement des fonctions continues en séries*, Math. Ann., 16 (1880), pp. 1-80.

[3] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, London, 1966, pp. 132-134.

[4] J. C. PIQUETTE, *An analytical expression for coefficients arising when implementing a technique for indefinite integration of products of special functions*, SIAM J. Math. Anal., 17 (1986), pp. 1033-1035.

[5] W. MAGNUS, F. OBERHETTINGER, AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, Berlin, New York, 1966, p. 191.

[6] Y. L. LUKE, *Integrals of Bessel Functions*, McGraw-Hill, New York, 1962.

[7] A. P. PRUDNIKOV, Y. A. BRYCHKOV, AND O. I. MARICHEV, *Integrals and Series*, Vol. 2 (N. M. Queen, trans.), Gordon and Breach, New York, 1986.

[8] J. PIQUETTE, *Uncoupling the differential equations arising from a technique for evaluating indefinite integrals containing special functions or their products*, submitted to Quart. Appl. Math.

SIAM J. MATH. ANAL.
Vol. 20, No. 5, pp. 1270–1282, September 1989

016

A NONTERMINATING q-CLAUSEN FORMULA AND SOME RELATED PRODUCT FORMULAS*

GEORGE GASPER† AND MIZAN RAHMAN‡

Abstract. This paper uses Gasper's proof of Rogers' linearization formula for the continuous q-ultraspherical polynomials and a quadratic transformation formula for well-poised basic hypergeometric ${}_2\phi_1$ series to derive a nonterminating q-analogue of Clausen's formula for the square of a certain hypergeometric series. This formula is extended to a q-analogue of the Ramanujan and Bailey extension of Clausen's formula by employing the Gasper and Rahman nonterminating q-extension of the Sears–Carlitz quadratic transformation formula. Additional product formulas are derived.

Key words. basic hypergeometric series, Clausen's formula, q-analogues, Ramanujan and Bailey product formula, nonnegative functions

AMS(MOS) subject classifications. primary 33A99; secondary 26D15

1. Introduction.

Clausen's [13] formula

$$\left\{{}_2F_1\left(a,b;a+b+\frac{1}{2}x\right)\right\}^2 = {}_3F_2\left(2a,2b,a+b;2a+2b,a+b+\frac{1}{2};x\right), \tag{1.1}$$

$|x|<1$, provides a rare example of a hypergeometric series that is expressible as the square of another hypergeometric series. Ramanujan's [25] rapidly convergent series representations for $1/\pi$, one of which was employed by Gosper in 1985 to compute π to more than 17,000,000 decimal digits, are based on special cases of (1.1); see the Chudnovskys' survey paper [12]. Clausen's formula was used in Askey and Gasper [4] to prove the nonnegativity of a certain sum of Jacobi polynomials which, in turn, played an important role in de Branges' [10] celebrated proof of the Bieberbach conjecture. One might say that it was de Branges' work that revived an interest in the methods that Askey and Gasper used in their proofs of the nonnegativity of certain sums and integrals of orthogonal polynomials; see, for example, Askey [1], [2] and Gasper [14].

Current literature in special functions also reveals a vigorous interest in generalizing almost every result in ordinary hypergeometric series to basic hypergeometric series. An ${}_r\phi_s$ basic hypergeometric series in base q is defined by

$${}_r\phi_s\left[\begin{matrix}a_1,\cdots,a_r\\ b_1,\cdots,b_s\end{matrix};q,z\right] = \sum_{n=0}^{\infty}\frac{(a_1,\cdots,a_r;q)_n}{(q,b_1,\cdots,b_s;q)_n}[(-1)^n q^{\binom{n}{2}}]^{1+s-r}z^n, \tag{1.2}$$

where $\binom{n}{2} = n(n-1)/2$,

$$(a_1,a_2,\cdots,a_r;q)_n = \prod_{j=1}^{r}(a_j;q)_n \tag{1.3}$$

---

*Received by the editors September 19, 1988; accepted for publication October 27, 1988.

†Department of Mathematics, Northwestern University, Evanston, Illinois 60208. The research of this author was supported in part by the National Science Foundation under grant DMS-8601901.

‡Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6, Canada. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant A6197.

and $(a;q)_n$ is the q-shifted factorial defined by

$$(1.4)\qquad (a;q)_n = \begin{cases} 1, & n=0 \\ (1-a)(1-aq)\cdots(1-aq^{n-1}), & n=1,2,\cdots \end{cases}.$$

The base q is usually a complex number with absolute value less than 1. The series (1.2) terminates if one of the numerator parameters is of the form $q^{-m}, m=0,1,2,\cdots$. It is assumed there are no zero factors in the denominators of the terms of the series. If $r=s+1$ and the series does not terminate then we assume that $|z|<1$ and $|q|<1$ to ensure convergence.

Jackson [21], who spent almost all of his mathematical career in studying basic hypergeometric series, derived the following product formula

$$(1.5)\qquad {}_2\phi_1\left[\begin{matrix} q^a, & q^b \\ & q^{a+b+1/2} \end{matrix}; q, x\right] {}_2\phi_1\left[\begin{matrix} q^a, & q^b \\ q^{(a+b)/2} & \end{matrix}; q, xq^{1/2}\right]$$
$$= {}_4\phi_3\left[\begin{matrix} q^a, q^b, & q^{(a+b)/2}, & -q^{(a+b)/2} & \\ & q^{a+b}, & q^{(a+b)/2+1/4}, & -q^{(a+b)/2+1/4} \end{matrix}; q, xq^{1/2}\right],$$

where $|x|<1$ and $|q|<1$, as a q-analogue of (1.1) in the sense that it tends to Clausen's formula as $q \to 1^-$. Additional proofs of (1.5) have been given by Singh [28], Nassrallah [23], and Jain and Srivastava [22].

Unfortunately, the left side of (1.5) is not a square and so (1.5) does not have this most important property of (1.1); in particular, (1.5) cannot be used to write certain sums of basic hypergeometric series as sums of squares of basic hypergeometric series as was done in [4], [14] for hypergeometric series. Recently, the authors independently derived (see [16]) such a formula by showing that

$$(1.6)\qquad \left\{{}_4\phi_3\left[\begin{matrix} a, & b, & abz, & ab/z \\ & abq^{1/2}, & -abq^{1/2}, & -ab \end{matrix}; q, q\right]\right\}^2$$
$$= {}_5\phi_4\left[\begin{matrix} a^2, & b^2, & ab, & abz, & ab/z \\ & a^2b^2, & abq^{1/2}, & -abq^{1/2}, & -ab \end{matrix}; q, q\right]$$

provided the series terminate. The terminating case of Clausen's formula follows from (1.6) by replacing a, b, z by $q^a, q^b, e^{i\theta}$, respectively, setting $x=(1-\cos\theta)/2$, and then letting $q \to 1$. Formula (1.6) was employed in [16] to prove the nonnegativity of certain basic hypergeometric series and to derive q-analogues of the Askey–Gasper [4] inequality

$$(1.7)\qquad \frac{(\alpha+2)_n}{n!}\,{}_3F_2\left[\begin{matrix} -n, & n+\alpha+2, & \frac{\alpha+1}{2} \\ & \frac{\alpha+3}{2}, & \alpha+1 \end{matrix}; x\right] \geq 0,$$

where $0 \leq x \leq 1, \alpha \geq -2$, and $n=0,1,2,\cdots$, and of the differential equations de Branges [10], [11] used in his proof of the Bieberbach conjecture.

Since (1.6) holds only when the series terminate (see [16]), while (1.1) holds for $|x|<1$ irrespective of whether the series terminate, it is irresistible to inquire whether there is a nonterminating q-Clausen formula that gives (1.6) in the terminating case, possibly after application of a transformation formula. As was pointed out in [16], there are several ways of deriving (1.6), such as using the Rogers' [27] linearization formula for the continuous q-ultraspherical polynomials $C_n(x;\beta|q)$, the Gasper and Rahman [17] formula for the product of terminating ${}_4\phi_3$ series, or the Rahman and

Verma [24] integral representation for the product $C_n(x;\beta|q)C_n(y;\beta|q)$. Even though each of these formulas involves only terminating series, we will show that a slight modification of Gasper's proof in [15] of Rogers' linearization formula leads to the following nonterminating q-analogue of Clausen's formula (1.1):

(1.8)

$$g^2(x) = \frac{\left(axq^{1/2}/b, bxq^{1/2}/a; q\right)_\infty}{\left(xq^{1/2}/ab, abxq^{1/2}; q\right)_\infty} {}_5\phi_4\left[\begin{matrix} a^2, b^2, ab, -ab, -abq^{1/2} \\ abq^{1/2}, a^2b^2, abxq^{1/2}, abq^{1/2}/x \end{matrix}; q, q\right]$$
$$+ \frac{(qx, qx, a^2, b^2; q)_\infty}{\left(abq^{1/2}, abq^{1/2}, ab/xq^{1/2}, abxq^{1/2}; q\right)_\infty}$$
$$\cdot {}_5\phi_4\left[\begin{matrix} axq^{1/2}/b, bxq^{1/2}/a, xq^{1/2}, -xq^{1/2}, -xq \\ xq, qx^2, abxq^{1/2}, xq^{3/2}/ab \end{matrix}; q, q\right],$$

where $(a;q)_\infty = \prod_{k=0}^{\infty}(1-aq^k)$, $\quad |q|<1$, $|qx|<1$,

$$g(x) = \frac{(qa^2x^2, qb^2x^2; q^2)_\infty}{(qx^2, qa^2b^2x^2; q^2)_\infty} {}_8W_7\left(-abxq^{-1/2}; a, b, -a, -b, -x; q, -qx\right), \tag{1.9}$$

and

(1.10)

$$ {}_{r+1}W_r(a; b_1, b_2, \cdots, b_{r-2}; q, z)$$
$$= {}_{r+1}\phi_r\left[\begin{matrix} a, qa^{1/2}, -qa^{1/2}, b_1, b_2, \cdots, b_{r-2} \\ a^{1/2}, -a^{1/2}, aq/b_1, aq/b_2, \cdots, aq/b_{r-2} \end{matrix}; q, z\right].$$

When $r = s+1$ the basic hypergeometric series (1.2) is said to be balanced if $z = q$ and $b_1b_2\cdots b_s = qa_1a_2\cdots a_{s+1}$; it is called well-poised if $b_1 = qa_1/a_2$, $b_2 = qa_1/a_3, \cdots, b_s = qa_1/a_{s+1}$, and it is called very well-poised if, in addition, $a_2 = qa_1^{1/2}, a_3 = -qa_1^{1/2}$. The ${}_4\phi_3$ series in (1.6) and the ${}_5\phi_4$ series in (1.6) and (1.8) are all balanced while the series in (1.9) and (1.10) are very well-poised.

We will derive (1.8) and some equivalent forms of it in §2 and show in §3 that the Clausen's formula (1.1) is a limit case of (1.8). Also, in §4 we shall derive a q-analogue of the Ramanujan [26] and Bailey [6], [7] extension of Clausen's formula

(1.11)

$$ {}_2F_1\left[\begin{matrix} a, b \\ c \end{matrix}; \frac{1-\sqrt{1-x}}{2}\right] {}_2F_1\left[\begin{matrix} a, b \\ a+b+1-c \end{matrix}; \frac{1-\sqrt{1-x}}{2}\right]$$
$$= {}_4F_3\left[\begin{matrix} a, & b, & (a+b)/2, & (a+b+1)/2 \\ & c, & a+b, & a+b+1-c \end{matrix}; x\right]$$

which gives (1.1) when $c = a+b+\frac{1}{2}$ and the quadratic transformation

$$ {}_2F_1\left[\begin{matrix} a, b \\ (a+b+1)/2 \end{matrix}; \frac{1-\sqrt{1-x}}{2}\right] {}_2F_1\left[\begin{matrix} a/2, b/2 \\ (a+b+1)/2 \end{matrix}; x\right] \tag{1.12}$$

is used. See Askey [3].

2. Proof of (1.8). Rogers' continuous q-ultraspherical polynomials, $C_n(x;\beta|q)$, are defined by

$$C_n(x;\beta|q) = \sum_{k=0}^{n} \frac{(\beta;q)_k(\beta;q)_{n-k}}{(q;q)_k(q;q)_{n-k}} \cos(n-2k)\theta$$

$$= \frac{(\beta;q)_n}{(q;q)_n} e^{in\theta}\, {}_2\phi_1\left[\begin{matrix} q^{-n}, & \beta \\ & \beta^{-1}q^{1-n} \end{matrix}; q, q\beta^{-1}e^{-2i\theta}\right] \tag{2.1}$$

where $x = \cos\theta$ and $n = 0, 1, \cdots$. Using an induction argument, Rogers [27, p. 29] proved that

$$C_m(x;\beta|q)C_n(x;\beta|q) = \sum_{k=0}^{\min(m,n)} a_k(m,n)C_{m+n-2k}(x;\beta|q), \tag{2.2}$$

where

$$a_k(m,n) = \frac{(\beta;q)_k(\beta;q)_{m-k}(\beta;q)_{n-k}(q;q)_{m+n-2k}}{(q;q)_k(q;q)_{m-k}(q;q)_{n-k}(\beta^2;q)_{m+n-2k}}$$

$$\cdot \frac{(\beta^2;q)_{m+n-k}(1-\beta q^{m+n-2k})}{(\beta q;q)_{m+n-k}(1-\beta)}. \tag{2.3}$$

Gasper [16] deduced (1.6) from (2.2) by setting $m = n$, using Askey and Ismail's [5] ${}_4\phi_3$ series representation for $C_n(x;\beta|q)$ and the ${}_6\phi_5$ summation formula [29, IV. 9].

The key to the discovery of a nonterminating q-Clausen formula is the observation that Gasper's [15] proof of (2.2) is independent of the fact that the parameter n in the ${}_2\phi_1$ series in (2.1) is a nonnegative integer.

In view of (2.1) let

$$f(x) = {}_2\phi_1\left[\begin{matrix} \alpha, & \beta \\ & \alpha q/\beta \end{matrix}; q, qx/\beta\right] \tag{2.4}$$

which reduces to the ${}_2\phi_1$ series in (2.1) when $\alpha = q^{-n}$ and $x = e^{-2i\theta}$. Temporarily assume that $|x| \le 1$, and $|q| < |\beta| < 1$. From Heine's transformation formula [8, §8.4, Eq. (2)] it follows that

$$f(x) = \frac{(\beta x;q)_\infty}{(qx/\beta;q)_\infty}\, {}_2\phi_1\left[\begin{matrix} \alpha q/\beta^2, & q/\beta \\ & \alpha q/\beta \end{matrix}; q, \beta x\right]. \tag{2.5}$$

Multiplying the two ${}_2\phi_1$ series in (2.4) and (2.5) and collecting the coefficients of x^j we find that

$$f^2(x) = \frac{(\beta x;q)_\infty}{(qx/\beta;q)_\infty} \sum_{j=0}^{\infty} A_j \frac{(\alpha q/\beta^2, q/\beta;q)_j}{(q, \alpha q/\beta;q)_j}(\beta x)^j, \tag{2.6}$$

where

$$A_j = {}_4\phi_3\left[\begin{matrix} q^{-j}, & \beta, & \beta q^{-j}/\alpha, & \alpha \\ & \beta q^{-j}, & \beta^2 q^{-j}/\alpha, & \alpha q/\beta \end{matrix}; q, q\right] \tag{2.7}$$

is a terminating balanced series. Following [15] we now use Watson's transformation formula [8, §8.5, Eq. (2)] to express the ${}_4\phi_3$ series in (2.7) as a very well-poised ${}_8\phi_7$ series:

$$A_j = \frac{(\alpha q/\beta, \alpha^2 q/\beta^2; q)_j}{(\alpha^2 q/\beta, \alpha q/\beta^2; q)_j}\, {}_8W_7\left(\alpha^2/\beta; \alpha, \alpha, \beta, \alpha^2 q^{j+1}/\beta^2, q^{-j}; q, q/\beta\right). \tag{2.8}$$

Using (2.8) in (2.6) immediately leads to the formula

$$\begin{aligned} f^2(x) = {} & \frac{(\beta x; q)_\infty}{(qx/\beta; q)_\infty} \sum_{k=0}^{\infty} \frac{(\alpha^2/\beta; q)_k (1 - \alpha^2 q^{2k}/\beta)(\alpha, \alpha, \beta; q)_k}{(q; q)_k (1 - \alpha^2/\beta)(\alpha q/\beta, \alpha q/\beta, \alpha^2 q/\beta^2; q)_k} \\ & \cdot \frac{(\alpha^2 q/\beta^2; q)_{2k}}{(\alpha^2 q/\beta; q)_{2k}} \left(\frac{qx}{\beta}\right)^k {}_2\phi_1\left[\begin{matrix} \alpha^2 q^{2k+1}/\beta^2, & q/\beta \\ & \alpha^2 q^{2k+1}/\beta \end{matrix}; q, \beta x\right]. \end{aligned} \tag{2.9}$$

The interesting property of the ${}_2\phi_1$ series in (2.4), (2.5), and (2.9) is that they are well poised and so we may apply the quadratic transformation formula [18, Eq. (3.8)]

$$\begin{aligned} & {}_2\phi_1\left[\begin{matrix} a, & b \\ & aq/b \end{matrix}; q, qx/b^2\right] \\ & = \frac{(qx/b, aqx^2/b^2; q)_\infty}{(aqx/b, qx^2/b^2; q)_\infty}\, {}_8W_7\left(ax/b; x, a^{1/2}, -a^{1/2}, (aq)^{1/2}, -(aq)^{1/2}; q, qx/b^2\right). \end{aligned} \tag{2.10}$$

We then use Bailey's transformation formula [8, §8.5, Eq. (3)] to express the resulting ${}_8W_7$ as a sum of two nonterminating balanced ${}_4\phi_3$ series. After some simplifications this gives

$$\begin{aligned} & {}_2\phi_1\left[\begin{matrix} \alpha^2 q^{2k+1}/\beta^2, & q/\beta \\ & \alpha^2 q^{2k+1}/\beta \end{matrix}; q, \beta x\right] \\ & = \frac{(\alpha x q/\beta; q)_\infty}{(\beta x/\alpha; q)_\infty} \frac{(-\alpha/\beta x)^k q^{\binom{k+1}{2}}}{(\alpha q/\beta x, \alpha x q/\beta; q)_k} \\ & \quad \cdot {}_4\phi_3\left[\begin{matrix} \alpha q^k, \alpha q^{k+1/2}/\beta, -\alpha q^{k+1/2}/\beta, -\alpha q^{k+1}/\beta \\ \alpha^2 q^{2k+1}/\beta, \alpha x q^{k+1}/\beta, \alpha q^{k+1}/\beta x \end{matrix}; q, q\right] \\ & \quad + \frac{(qx, \alpha, \alpha x q, \alpha^2 q/\beta^2; q)_\infty}{(\beta x, \alpha q/\beta, \alpha/\beta x, \alpha^2 q/\beta; q)_\infty} \cdot \frac{(\alpha q/\beta, \alpha/\beta x; q)_k}{(\alpha, \alpha x q; q)_k} \frac{(\alpha^2 q/\beta; q)_{2k}}{(\alpha^2 q/\beta^2; q)_{2k}} \\ & \quad \cdot {}_4\phi_3\left[\begin{matrix} \beta x, x q^{1/2}, -x q^{1/2}, -xq \\ qx^2, \alpha x q^{k+1}, \beta x q^{1-k}/\alpha \end{matrix}; q, q\right]. \end{aligned} \tag{2.11}$$

Since we have assumed that $|x| \le 1$ and $|q| < |\beta| < 1$, we can now substitute (2.11) into (2.9) and change the order of summation to obtain the formula

$$\begin{aligned} f^2(x) = {} & \frac{(\beta x, \alpha x q/\beta; q)_\infty}{(qx/\beta, \beta x/\alpha; q)_\infty} \sum_{m=0}^{\infty} \frac{(\alpha; q)_m (\alpha^2 q/\beta^2; q)_{2m} q^m}{(q, \alpha x q/\beta, \alpha q/\beta x, \alpha^2 q/\beta, \alpha q/\beta; q)_m} \\ & \cdot {}_6W_5\left(a^2/\beta; a, \beta, q^{-m}; q, \alpha q^{m+1}/\beta^2\right) \\ & + \frac{(\alpha, \alpha^2 q/\beta^2, qx, \alpha q x; q)_\infty}{(\alpha q/\beta, \alpha^2 q/\beta, qx/\beta, \alpha/\beta x; q)_\infty} \sum_{m=0}^{\infty} \frac{(\beta x, x q^{1/2}, -x q^{1/2}, -xq; q)_m}{(q, qx^2, \alpha x q, \beta x q/\alpha; q)_m} q^m \\ & \cdot {}_6W_5\left(\alpha^2/\beta; \alpha, \beta, \alpha q^{-m}/\beta x; q, x q^{m+1}/\beta\right). \end{aligned} \tag{2.12}$$

By [29, IV. 9]

$$(2.13)\qquad {}_6W_5\left(\alpha^2/\beta;\alpha,\beta,q^{-m};q,\alpha q^{m+1}/\beta^2\right)=\frac{(\alpha^2q/\beta,\alpha q/\beta^2;q)_m}{(\alpha q/\beta,\alpha^2q/\beta^2;q)_m},$$

and by [29, IV. 7]

$$(2.14)\qquad \begin{aligned}&{}_6W_5\left(a^2/\beta;\alpha,\beta,\alpha q^{-m}/\beta x;q,xq^{m+1}/\beta\right)\\&=\frac{(\alpha^2q/\beta,\alpha q/\beta^2,xq,\alpha xq/\beta;q)_\infty}{(\alpha q/\beta,\alpha^2q/\beta^2,\alpha xq,xq/\beta;q)_\infty}\frac{(\alpha xq,xq/\beta;q)_m}{(xq,\alpha xq/\beta;q)_m}.\end{aligned}$$

Using (2.13) and (2.14) we then obtain the formula

$$(2.15)\qquad \begin{aligned}&\left\{{}_2\phi_1\left[\begin{matrix}\alpha,\beta\\ q/\beta\end{matrix};q,qx/\beta\right]\right\}^2\\&=\frac{(\beta x,\alpha xq/\beta;q)_\infty}{(xq/\beta,\beta x/\alpha;q)_\infty}\,{}_5\phi_4\left[\begin{matrix}\alpha,\alpha q^{1/2}/\beta,\alpha q/\beta^2,-\alpha q^{1/2}/\beta,-\alpha q/\beta\\ \alpha q/\beta,\alpha^2q/\beta^2,\alpha q/\beta x,\alpha xq/\beta\end{matrix};q,q\right]\\&\quad+\frac{(\alpha,\alpha q/\beta^2,xq,xq,\alpha xq/\beta;q)_\infty}{(\alpha q/\beta,\alpha q/\beta,xq/\beta,xq/\beta,\alpha/\beta x;q)_\infty}\\&\quad\cdot{}_5\phi_4\left[\begin{matrix}\beta x,&xq/\beta,&xq^{1/2},&-xq^{1/2},&-xq\\ &qx,&qx^2,&\beta xq/\alpha,&\alpha xq/\beta\end{matrix};q,q\right],\end{aligned}$$

which gives the square of a well-poised ${}_2\phi_1$ series as the sum of two balanced ${}_5\phi_4$ series. By analytic continuation, (2.15) holds when $|q|<1$ and $|qx/\beta|<1$.

If $\alpha=q^{-n},n=0,1,\cdots$, then $(\alpha;q)_\infty=0$ and so

$$(2.16)\qquad \begin{aligned}f^2(x)&=\frac{(\beta x,xq^{1-n}/\beta;q)_\infty}{(\beta xq^n,qx/\beta;q)_\infty}\\&\quad\cdot{}_5\phi_4\left[\begin{matrix}q^{-n},&q^{1/2-n}/\beta,&q^{1-n}/\beta^2,&-q^{1/2-n}/\beta,&-q^{1-n}/\beta\\ &q^{1-n}/\beta,&q^{1-2n}/\beta^2,&xq^{1-n}/\beta,&q^{1-n}/\beta x\end{matrix};q,q\right]\\&=b_n\,{}_5\phi_4\left[\begin{matrix}q^{-n},&\beta^2q^n,&\beta,&\beta x,&\beta/x\\ &\beta^2,&\beta q^{1/2},&-\beta q^{1/2},&-\beta\end{matrix};q,q\right],\end{aligned}$$

where

$$(2.17)\qquad \begin{aligned}b_n&=\frac{(\beta x,q^{-n},q^{1-n}/\beta^2,q^{1/2-n}/\beta,-q^{1/2-n}/\beta,-q^{1-n}/\beta;q)_n}{(q,q^{1-n}/\beta,q^{1-2n}/\beta^2,q^{1-n}/\beta x;q)_n}q^n\\&=\frac{(\beta^2,\beta q^{1/2},-\beta q^{1/2},-\beta;q)_n}{(\beta,\beta^2q^n;q)_n}\left(\frac{x}{\beta}\right)^n=\frac{(\beta^2,\beta^2;q)_n}{(\beta,\beta;q)_n}\left(\frac{x}{\beta}\right)^n.\end{aligned}$$

However, by [5, Eq. (3.10)],

$$(2.18)\qquad f(x)=\frac{(\beta^2;q)_n}{(\beta;q)_n}\left(\frac{x}{\beta}\right)^{n/2}{}_4\phi_3\left[\begin{matrix}q^{-n},&\beta^2q^n,&(\beta x)^{1/2},&(\beta/x)^{1/2}\\ &\beta q^{1/2},&-\beta q^{1/2},&-\beta\end{matrix};q,q\right].$$

It follows from (2.16)–(2.18) that

$$(2.19)\qquad \begin{aligned}&\left\{{}_4\phi_3\left[\begin{matrix}q^{-n},&\beta^2q^n,&(\beta x)^{1/2},&(\beta/x)^{1/2}\\ &\beta q^{1/2},&-\beta q^{1/2},&-\beta\end{matrix};q,q\right]\right\}^2\\&={}_5\phi_4\left[\begin{matrix}q^{-n},&\beta^2q^n,&\beta,&\beta x,&\beta/x\\ &\beta^2,&\beta q^{1/2},&-\beta q^{1/2},&-\beta\end{matrix};q,q\right],\end{aligned}$$

which is formula (1.6) written in a special form.

Before closing this section we need to show that (2.15) is equivalent to (1.8). By (2.10) and Bailey's ${}_8\phi_7$ transformation formula [9, Eq. (4.3)], we have

$$
\begin{aligned}
&{}_2\phi_1\left[\begin{matrix}\alpha, & \beta\\ & \alpha q/\beta\end{matrix}; q, qx/\beta\right]\\
&\quad = \frac{(x(\alpha q)^{1/2}, -x(\alpha q)^{1/2}, qx\alpha^{1/2}/\beta, -qx\alpha^{1/2}/\beta; q)_\infty}{(xq^{1/2}, -xq^{1/2}, qx/\beta, -\alpha xq/\beta; q)_\infty}\\
&\quad\cdot {}_8W_7\left(-\alpha x/\beta; \alpha^{1/2}, (\alpha q)^{1/2}/\beta, -\alpha^{1/2}, -(\alpha q)^{1/2}/\beta, -x; q, -qx\right).
\end{aligned}
\tag{2.20}
$$

Now set $a = \alpha^{1/2}$ and $b = (\alpha q)^{1/2}/\beta$ to obtain from (2.15) that

$$
\begin{aligned}
&\left\{\frac{(qa^2x^2, qb^2x^2; q^2)_\infty}{(qx^2, qa^2b^2x^2; q^2)_\infty}\, {}_8W_7\left(-qbxq^{-1/2}; a, b, -a, -b, -x; q, -qx\right)\right\}^2\\
&\quad = \frac{(axq^{1/2}/b, bxq^{1/2}/a; q)_\infty}{(xq^{1/2}/ab, abxq^{1/2}; q)_\infty}\, {}_5\phi_4\left[\begin{matrix}a^2, b^2, ab, -ab, -abq^{1/2}\\ abq^{1/2}, a^2b^2, abxq^{1/2}, abq^{1/2}/x\end{matrix}; q, q\right]\\
&\qquad + \frac{(qx, qx, a^2, b^2; q)_\infty}{(abq^{1/2}, abq^{1/2}, ab/xq^{1/2}, abxq^{1/2}; q)_\infty}\\
&\qquad\cdot {}_5\phi_4\left[\begin{matrix}axq^{1/2}/b, bxq^{1/2}/a, xq^{1/2}, -xq^{1/2}, -xq\\ qx, qx^2, abxq^{1/2}, xq^{3/2}/ab\end{matrix}; q, q\right], \quad |qx|<1, \quad |q|<1,
\end{aligned}
\tag{2.21}
$$

which is formula (1.8).

3. Some variations of (2.9). If we replace a and b by q^a and q^b, respectively, in (2.21) and take the limit $q \to 1^-$, then the left side can be seen to approach the left side of (1.1) with x replaced by $-4x(1-x)^{-2}$ while the first term on the right of (2.21) approaches the right side of (1.1) with x replaced by $-4x(1-x)^{-2}$. So (2.21) can justifiably be called a q-Clausen formula provided we can show that the limit of the second term on the right side as $q \to 1^-$ is zero. This can be done by using the asymptotics of the theta functions somewhat along the lines followed in [24], but an easier approach is to use (2.20) for the ${}_2\phi_1$ function in (2.9) and then take the term-by-term limit from a uniformly convergent double series. First, by an obvious transformation of summation indices we obtain from (2.9), (2.20), and (1.9) that

$$
\begin{aligned}
g^2(x) &= \frac{(a^2xq^{1/2}, -a^2xq^{1/2}, -abxq^{1/2}, bxq^{1/2}/a; q)_\infty}{(xq^{1/2}, -xq^{1/2}, abxq^{1/2}, -a^3bxq^{1/2}; q)_\infty}\\
&\cdot \sum_{m=0}^{\infty} \frac{(-a^3bxq^{-1/2}; q)_m(1+a^3bxq^{2m-1/2})(a^2, ab, -a^2, -ab, -x; q)_m}{(q; q)_m(1+a^3bxq^{-1/2})(-abxq^{1/2}, -a^2xq^{1/2}, abxq^{1/2}, a^2xq^{1/2}, a^3bq^{1/2}; q)_m}\\
&\cdot (-xq)^m\, {}_8W_7\left(a^3bq^{-1/2}; a^2, aq^{1/2}/b, -abq^{1/2}, -a^3bxq^{m-1/2}, q^{-m}; q, bq^{1/2}/ax\right).
\end{aligned}
\tag{3.1}
$$

Then, replacing a, b by q^a, q^b and letting $q \to 1^-$, we obtain

$$
\left\{{}_2F_1(a, b; a+b+\tfrac{1}{2}; z)\right\}^2 = \sum_{m=0}^{\infty} \frac{(2a)_m(a+b)_m}{m!(3a+b+\frac{1}{2})_m} z^m
\tag{3.2}
$$

$$\cdot {}_5F_4\left[\begin{matrix} 3a+b-\frac{1}{2}, (3a+b)/2+\frac{3}{4}, 2a, a-b+\frac{1}{2}, -m \\ (3a+b/2)-\frac{1}{4}, a+b+\frac{1}{2}, 2a+2b, 3a+b+\frac{1}{2}+m \end{matrix}; 1\right]$$

$$= \sum_{m=0}^{\infty} \frac{(2a)_m (a+b)_m}{m!(3a+b+\frac{1}{2})_m} z^m \frac{(3a+b+\frac{1}{2})_m (2b)_m}{(a+b+\frac{1}{2})_m (2a+2b)_m} \qquad \text{(by [29, III.13])}$$

$$= {}_3F_2\left[\begin{matrix} 2a, & 2b, & a+b \\ & a+b+\frac{1}{2}, & 2a+2b \end{matrix}; z\right]$$

where $z = -4x(1-x)^{-2}$. This completes the proof that (2.21) is a q-analogue of Clausen's formula (1.1). Clausen's formula can be used to show that if a and b are replaced by q^a and q^b in (2.21), then the second term on the right side tends to zero as $q \to 1^-$.

Similarly, it can be shown that formula (2.15) is a q-analogue of the formula

$$\left\{ {}_2F_1\left[\begin{matrix} a, & b \\ a+1-b \end{matrix}; x\right]\right\}^2$$

(3.3)

$$= (1-x)^{-2a} \, {}_3F_2\left[\begin{matrix} a, a+\frac{1}{2}-b, a+1-2b \\ a+1-b, 2a+1-2b \end{matrix}; -\frac{4x}{(1-x)^2}\right],$$

where $|x| < 1$ and $|4x(1-x)^{-2}| < 1$.

It may be of interest to point out some other aspects of (3.1). The very well-poised ${}_8W_7$ series on the right side of (3.1) is terminating and so can be transformed to a terminating balanced ${}_4\phi_3$ series by Watson's formula [8, §8.5, Eq. (2)]. Thus we find that

$$g^2(x) = \frac{(a^2xq^{1/2}, -a^2xq^{1/2}, -abxq^{1/2}, bxq^{1/2}/a; q)_\infty}{(xq^{1/2}, -xq^{1/2}, abxq^{1/2}, -a^3bxq^{1/2}; q)_\infty}$$

$$\cdot \sum_{m=0}^{\infty} \frac{(-a^3bxq^{-1/2}; q)_m (1+a^3bxq^{2m-1/2})(a^2, ab, -ab; q)_m}{(q; q)_m (1+a^3bxq^{-1/2})(-abxq^{1/2}, -a^2xq^{1/2}, a^2xq^{1/2}; q)_m} \tag{3.4}$$

$$\cdot \left(\frac{xq^{1/2}}{ab}\right)^m {}_4\phi_3\left[\begin{matrix} q^{-m}, & -a^3bxq^{m-1/2}, & b^2, & -abq^{1/2} \\ & a^2b^2, & abq^{1/2}, & abxq^{1/2} \end{matrix}; q, q\right].$$

If a^2 or ab or $-ab$ is of the form q^{-n}, $n = 0, 1, 2, \cdots$, then by changing the order of summation the double sum on the right side becomes a sum of very well-poised ${}_6\phi_5$ series which can be summed by [29, IV. 9], reducing the right side of (3.4) to a multiple of a ${}_5\phi_4$ series. This is essentially the route taken in [16] to derive (1.6). In general, this change in order of summation is not valid since the double series does not converge absolutely, as can be seen by observing that if it were absolutely convergent then we would have

$$g^2(x) = \frac{(a^2xq^{1/2}, -a^2xq^{1/2}, -abxq^{1/2}, bxq^{1/2}/a; q)_\infty}{(xq^{1/2}, -xq^{1/2}, abxq^{1/2}, -a^3bxq^{1/2}; q)_\infty}$$

$$\cdot \sum_{k=0}^{\infty} \frac{(a^2, b^2, ab, -ab, -abq^{1/2}; q)_k (-a^3bxq^{1/2}; q)_{2k}}{(q, a^2b^2, abq^{1/2}, abxq^{1/2}, -abxq^{1/2}, a^2xq^{1/2}, -a^2xq^{1/2}; q)_k} \tag{3.5}$$

$$\cdot q^{-k^2/2} \left(-\frac{xq}{ab}\right)^k {}_6W_5\left(-a^3bxq^{2k-1/2}; a^2q^k, abq^k, -abq^k; q, xq^{1/2-k}/ab\right),$$

but the ${}_6W_5$ series clearly diverges for sufficiently large values of k when $x \neq 0$.

By applying Bailey's [9, Eqs. (4.3)–(4.6)] transformation formulas to the ${}_8W_7$ series in (2.21) and transformation formulas for well-poised series to the ${}_2\phi_1$ series in (2.15), these formulas can be written in many equivalent forms. In particular, since it follows from (3.8) and (3.9) in [18] that

$$(3.6)\qquad {}_4\phi_3\left[\begin{matrix} a, & qa^{1/2}, & -qa^{1/2}, & b \\ & a^{1/2}, & -a^{1/2}, & aq/b \end{matrix}; q, x\right] = (1-bx)\,{}_2\phi_1\left[\begin{matrix} aq, bq \\ aq/b \end{matrix}; q, x\right],$$

which can also be verified directly, (2.15) can be used to write the square of a very well-poised ${}_4\phi_3$ series as a sum of two balanced ${}_5\phi_4$ series. In addition, Jackson's [20] transformation formula

$$(3.7)\qquad {}_2\phi_1\left[\begin{matrix} a, & b \\ & c \end{matrix}; q, x\right] = \frac{(ax;q)_\infty}{(x;q)_\infty}\,{}_2\phi_2\left[\begin{matrix} a, & c/b \\ c, & ax \end{matrix}; q, bx\right]$$

can be applied to the ${}_2\phi_1$ series in (2.15) to derive formulas which after changes in variables are q-analogues of the formulas

$$(3.8)\qquad \left\{{}_2F_2\left[\begin{matrix} a, & b \\ (a+b+1)/2 \end{matrix}; x\right]\right\}^2 = {}_3F_2\left[\begin{matrix} a, b, (a+b)/2 \\ a+b, & (a+b+1)/2 \end{matrix}; 4x(1-x)\right]$$

and

$$(3.9)\qquad \left\{{}_2F_1\left[\begin{matrix} a, & 1-a \\ & c \end{matrix}; x\right]\right\}^2 = (1-x)^{2c-2}\,{}_3F_2\left[\begin{matrix} a+c-1, c-a, c-\frac{1}{2} \\ c, 2c-1 \end{matrix}; 4x(1-x)\right],$$

where $|x|<1$ and $|4x(1-x)|<1$.

4. Additional product formulas. Let us start by proceeding as in the derivation of (2.6) to derive the more general formula

$$\begin{aligned} &{}_2\phi_1\left[\begin{matrix} a, & b \\ & c \end{matrix}; q, x\right]\,{}_2\phi_1\left[\begin{matrix} \alpha, & \beta \\ & \gamma \end{matrix}; q, y\right] \\ &\quad = \frac{(\alpha\beta y/\gamma;q)_\infty}{(y;q)_\infty}\,{}_2\phi_1\left[\begin{matrix} a, & b \\ & c \end{matrix}; q, x\right]\,{}_2\phi_1\left[\begin{matrix} \gamma/\alpha, & \gamma/\beta \\ & \gamma \end{matrix}; q, \frac{\alpha\beta y}{\gamma}\right] \\ &\quad = \frac{(\alpha\beta y/\gamma;q)_\infty}{(y;q)_\infty}\sum_{j=0}^{\infty} B_j \frac{(\gamma/\alpha, \gamma/\beta; q)_j}{(q, \gamma; q)_j}\left(\frac{\alpha\beta y}{\gamma}\right)^j \end{aligned} \qquad (4.1)$$

with

$$(4.2)\qquad B_j = {}_4\phi_3\left[\begin{matrix} q^{-j}, & q^{1-j}/\gamma, & a, & b \\ & \alpha q^{1-j}/\gamma, & \beta q^{1-j}/\gamma, & c \end{matrix}; q, \frac{xq}{y}\right]$$

provided $\max(|x|, |y|, |\alpha\beta y/\gamma|) < 1$. In order to apply Watson's transformation formula [8, §8.5, Eq. (2)] to the ${}_4\phi_3$ series in (4.2), it is necessary that the series be balanced, that is

$$(4.3)\qquad \gamma = \alpha\beta c/ab \qquad \text{and} \qquad y = x.$$

Then, assuming that (4.3) holds,

$$B_j = \frac{(\alpha c/a, \alpha c/b; q)_j}{(\alpha c, \alpha c/ab; q)_j}\, {}_8W_7\left(\alpha c/q; a, b, \alpha, \alpha\beta c^2 q^{j-1}/ab, q^{-j}; q, q/\beta\right) \tag{4.4}$$

and hence

$$\begin{aligned}
&{}_2\phi_1\left[\begin{matrix} a, & b \\ & c \end{matrix}; q, x\right]\ {}_2\phi_1\left[\begin{matrix} \alpha, & \beta \\ & \alpha\beta c/ab \end{matrix}; q, x\right] \\
&= \frac{(abx/c; q)_\infty}{(x; q)_\infty} \sum_{j=0}^{\infty} \frac{(\beta c/ab, \alpha c/a, \alpha c/b; q)_j}{(q, \alpha\beta c/ab, \alpha c; q)_j} \left(\frac{abx}{c}\right)^j \\
&\quad \cdot \sum_{k=0}^{j} \frac{(1-\alpha c q^{2k-1})(\alpha c/q, a, b, \alpha, \alpha\beta c^2 q^{j-1}/ab, q^{-j}; q)_k}{(1-\alpha c/q)(q, \alpha c/a, \alpha c/b, c, abq^{1-j}/\beta c, \alpha c q^j; q)_k} \left(\frac{q}{\beta}\right)^k \\
&= \frac{(abx/c; q)_\infty}{(x; q)_\infty} \sum_{k=0}^{\infty} C_k \frac{(1-\alpha c q^{2k-1})\,(\alpha c/q, a, b, \alpha; q)_k (\alpha\beta c^2/abq; q)_{2k}}{(1-\alpha c/q)(q, c, \alpha\beta c/ab, \alpha\beta c^2/abq; q)_k (\alpha c; q)_{2k}} x^k
\end{aligned} \tag{4.5}$$

with

$$C_k = {}_4\phi_3\left[\begin{matrix} \alpha\beta c^2 q^{2k-1}/ab, & \beta c/ab, & \alpha c q^k/a, & \alpha c q^k/b \\ & \alpha c q^{2k}, & \alpha\beta c q^k/ab, & \alpha\beta c^2 q^{k-1}/ab \end{matrix}; q, \frac{abx}{c}\right]. \tag{4.6}$$

To proceed further we now observe that the quadratic transformation formula (2.11) is a special case of the authors' nonterminating extension [18, Eq. (1.2)]

$$\begin{aligned}
&{}_3\phi_2\left[\begin{matrix} a, & b, & c \\ & aq/b, & aq/c \end{matrix}; q, \frac{aqx}{bc}\right] \\
&= \frac{(ax; q)_\infty}{(x; q)_\infty}\, {}_5\phi_4\left[\begin{matrix} a^{1/2}, & -a^{1/2}, & (aq)^{1/2}, & -(aq)^{1/2}, & aq/bc \\ & aq/b, & aq/c, & ax, & q/x \end{matrix}; q, q\right] \\
&\quad + \frac{(a, aq/bc, aqx/b, aqx/c; q)_\infty}{(aq/b, aq/c, aqx/bc, x^{-1}; q)_\infty} \\
&\quad \cdot {}_5\phi_4\left[\begin{matrix} xa^{1/2}, & -xa^{1/2}, & x(aq)^{1/2}, & -x(aq)^{1/2}, & aqx/bc \\ & aqx/b, & aqx/c, & xq, & ax^2 \end{matrix}; q, q\right]
\end{aligned} \tag{4.7}$$

of the Sears–Carlitz transformation formula for a terminating well-poised ${}_3\phi_2$ series. Thus, if we set $\alpha = a$ and $\beta = aq/c$ then the ${}_4\phi_3$ series in (4.6) reduces to a well-poised ${}_3\phi_2$ series to which we can apply (4.7) to extend (2.11) to

$$\begin{aligned}
&{}_3\phi_2\left[\begin{matrix} acq^{2k}/b, q/b, cq^k \\ acq^{2k}, aq^{k+1}/b \end{matrix}; q, \frac{abx}{c}\right] \\
&= \frac{(ax; q)_\infty (-cq/bx)^k q^{\binom{k}{2}}}{(bx/c; q)_\infty (ax, cq/bx; q)_k} \\
&\quad \cdot {}_5\phi_4\left[\begin{matrix} aq^k, (ac/b)^{1/2}q^k, -(ac/b)^{1/2}q^k, (ac/b)^{1/2}q^{k+1/2}, -(ac/b)^{1/2}q^{k+1/2} \\ & acq^{2k}, aq^{k+1}/b, axq^k, cq^{k+1}/bx \end{matrix}; q, q\right] \\
&\quad + \frac{(ac/b, a, abx, aqx/c; q)_\infty (aq/b, c/bx; q)_k (ac; q)_{2k}}{(ac, aq/b, abx/c, c/bx; q)_\infty (a, abx; q)_k (ac/b; q)_{2k}} \\
&\quad \cdot {}_5\phi_4\left[\begin{matrix} x(ab/c)^{1/2}, -x(ab/c)^{1/2}, x(abq/c)^{1/2}, -x(abq/c)^{1/2}, abx/c \\ & abxq^k, aqx/c, bxq^{1-k}/c, abx^2/c \end{matrix}; q, q\right].
\end{aligned} \tag{4.8}$$

Substituting (4.8) into (4.5) and changing the order of summation, we find that

$$
\begin{aligned}
&{}_2\phi_1\left[\begin{matrix}a, & b\\ & c\end{matrix};q,x\right]\ {}_2\phi_1\left[\begin{matrix}a, & aq/c\\ & aq/b\end{matrix};q,x;q,x\right]\\
&\quad=\frac{(ax,abx/c;q)_\infty}{(x,bx/c;q)_\infty}\sum_{m=0}^{\infty}\frac{(a;q)_m(ac/b;q)_{2m}q^m}{(q,aq/b,ax,cq/bx,ac;q)_m}\\
&\quad\quad\cdot\ {}_6W_5\left(ac/q;a,b,q^{-m};q,cq^m/b\right)\\
&\quad\quad+\frac{(ac/b,a,abx,aqx/c;q)_\infty}{(ac,aq/b,x,c/bx;q)_\infty}\sum_{m=0}^{\infty}\frac{(abx/c;q)_m(abx^2/c;q)_{2m}q^m}{(q,abx^2/c,aqx/c,abx,bxq/c;q)_m}\\
&\quad\quad\cdot\ {}_6W_5\left(ac/q;a,b,cq^{-m}/bx;q,xq^m\right).
\end{aligned}
\tag{4.9}
$$

Since each of the above ${}_6W_5$ series is summable by [29, IV. 7], it follows from (4.9) that we have the product formula

$$
\begin{aligned}
&{}_2\phi_1\left[\begin{matrix}a, & b\\ & c\end{matrix};q,x\right]\ {}_2\phi_1\left[\begin{matrix}a, & aq/c\\ & aq/b\end{matrix};q,x\right]\\
&\quad=\frac{(ax,abx/c;q)_\infty}{(x,bx/c;q)_\infty}\ {}_6\phi_5\left[\begin{matrix}a,c/b,(ac/b)^{1/2},-(ac/b)^{1/2},(acq/b)^{1/2},-(acq/b)^{1/2}\\ aq/b,c,ac/b,ax,cq/bx\end{matrix};q,q\right]\\
&\quad\quad+\frac{(a,c/b,ax,bx,axq/c;q)_\infty}{(c,aq/b,x,x,c/bx;q)_\infty}\\
&\quad\quad\cdot\ {}_6\phi_5\left[\begin{matrix}x,abx/c,x(ab/c)^{1/2},-x(ab/c)^{1/2},x(abq/c)^{1/2},-x(abq/c)^{1/2}\\ ax,bx,axq/c,bxq/c,abx^2/c\end{matrix};q,q\right],
\end{aligned}
\tag{4.10}
$$

where $|x|<1$ and $|q|<1$. This formula reduces to (2.15) when $a=\alpha, b=\beta, c=\alpha q/\beta$ and x is replaced by qx/β.

By applying transformation formulas to the ${}_2\phi_1$ series in (4.10), this formula can be written in many equivalent forms. In particular, to derive a q-analogue of the Ramanujan and Bailey product formula (1.11), replace b in (4.10) by c/b and apply Jackson's transformation formula (3.7) to obtain

$$
\begin{aligned}
&{}_2\phi_2\left[\begin{matrix}a, & b\\ c, & ax\end{matrix};q,\frac{cx}{b}\right]\ {}_2\phi_2\left[\begin{matrix}a, & b\\ abq/c, & ax\end{matrix};q,\frac{axq}{c}\right]\\
&\quad=\frac{(x,ax/b;q)_\infty}{(ax,x/b;q)_\infty}\ {}_6\phi_5\left[\begin{matrix}a,b,(ab)^{1/2},-(ab)^{1/2},(abq)^{1/2},-(abq)^{1/2}\\ abq/c,c,ab,ax,bq/x\end{matrix};q,q\right]\\
&\quad\quad+\frac{(a,b,cx/b,axq/c;q)_\infty}{(c,abq/c,ax,b/x;q)_\infty}\\
&\quad\quad\cdot\ {}_6\phi_5\left[\begin{matrix}x,abx,x(a/b)^{1/2},-x(a/b)^{1/2},x(aq/b)^{1/2},-x(aq/b)^{1/2}\\ ax,cx/b,axq/c,xq/b,ax^2/b\end{matrix};q,q\right]
\end{aligned}
\tag{4.11}
$$

when $\max(|q|,|axq/c|,|cx/b|)<1$. If we replace a,b,x in (4.11) by $q^a,q^b,z/(z-1)$, respectively, and let $q\to 1^-$, we get

$$
\begin{aligned}
&{}_2F_1\left[\begin{matrix}a, & b\\ & c\end{matrix};z\right]\ {}_2F_1\left[\begin{matrix}a, & b\\ & a+b-c+1\end{matrix};z\right]\\
&\quad={}_4F_3\left[\begin{matrix}a, & b, & (a+b)/2, & (a+b+1)/2\\ & c, & a+b, & a+b+1-c\end{matrix};4z(1-z)\right],
\end{aligned}
\tag{4.12}
$$

which on setting $z = (1 - \sqrt{1-x})/2$ gives (1.11). Note that when $c = (abq)^{1/2}$, formula (4.11) gives the square of the series

$$ {}_2\phi_2\left[\begin{matrix} a, & b \\ (abq)^{1/2}, & ax \end{matrix}; q, x(aq/b)^{1/2}\right] $$

as a sum of two balanced ${}_5\phi_4$ series.

REFERENCES

[1] R. ASKEY, *Orthogonal polynomials and positivity*, Special Functions and Wave Propagation, D. Ludwig and F. W. J. Oliver, eds., Society for Industrial and Applied Mathematics, Philadelphia (1970), pp. 64-85.

[2] ———, *Orthogonal Polynomials and Special Functions*, CBMS–NSF Regional Conference Series in Applied Mathematics, 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[3] ———, *Variants of Clausen's formula for the square of a special* ${}_2F_1$, to appear.

[4] R. ASKEY AND G. GASPER, *Positive Jacobi polynomial sums*, II. Amer. J. Math., 98 (1976), pp. 709-737.

[5] R. ASKEY AND M. E. H. ISMAIL, *A generalization of ultraspherical polynomials*, in Studies in Pure Mathematics, P. Erdös, ed., Birkhäuser, Basel, Switzerland (1983), pp. 55-78.

[6] W. N. BAILEY, *A reducible case of the fourth type of Appell's hypergeometric functions of two variables*, Quart. J. Math. (Oxford), 4 (1933), pp. 305-308.

[7] ———, *Some theorems concerning products of hypergeometric series*, Proc. Lond. Math. Sci. 38 (1935), pp. 377-384.

[8] ———, Generalized Hypergeometric Series, Cambridge University Press, Cambridge, 1935; reprinted by Stechert–Hafner Agency, New York, London, 1964.

[9] ———, *Series of hypergeometric type which are infinite in both directions*, Quart. J. Math. (Oxford), 7 (1936), pp. 105-115.

[10] L. DE BRANGES, A proof of the Bieberbach conjecture, Acta Math., 154 (1985), pp. 137-152.

[11] ———, *Powers of Riemann mapping functions*, in The Bieberbach Conjecture: Proc. of the Symposium on the Occasion of the Proof, A. Baernstein, D. Drasin, P. Duren, and A. Marden, eds., Math. Surveys and Monographs, 21, 1986, American Mathematical Society, Providence, R.I., pp. 51-67.

[12] D. V. CHUDNOVSKY AND G. V. CHUDNOVSKY, *Approximations and complex multiplication according to Ramanujan*, in Ramanujan Revisited, Proceedings of the Centenary Conference, G.E. Andrews, R. Askey, B.C. Berndt, K.G. Ramanathan, and R.A. Rankin, eds., Academic Press, New York, 1988, pp. 375-472.

[13] T. CLAUSEN, *Ueber die Fälle, wenn die Reihe von der Form* $y = 1+x\alpha\beta/1.\gamma + \cdots$ *ein Quadrat von der Form* $z = 1 + x\alpha'\beta'\gamma'/1.\delta'\epsilon' + \cdots$ *hat*, J. für Math., 3 (1828), pp. 89-91.

[14] G. GASPER, *Positivity and special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 375-433.

[15] ———, *Rogers' linearization formula for the continuous q-ultraspherical polynomials and quadratic transformation formulas*, SIAM J. Math. Anal., 16 (1985), pp. 1061-1071.

[16] ———, q-Extensions of Clausen's formula and of the inequalities used by de Branges in his proof of the Bieberbach, Robertson, and Milin conjectures, 20 (1989), pp. 1019–1034.

[17] G. GASPER AND M. RAHMAN, *Product formulas of Watson, Bailey and Bateman types and positivity of the Poisson kernel for q-Racah polynomials*, SIAM J. Math. Anal., 15 (1984), pp. 768-789.

[18] ———, *Positivity of the Poisson kernal for the continuous q-Jacobi polynomials and some quadratic transformation formulas for basic hypergeometric series*, SIAM J. Math. Anal., 17 (1986), pp. 970-999.

[19] ———, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, to appear.

[20] F. H. JACKSON, *Transformation of q-series*, Messenger of Math., 39 (1910), pp. 145-153.

[21] ———, *The* q^θ *equations whose solutions are products of solutions of* q^θ *equations of lower order*, Quart. J. Math. (Oxford), 11 (1940), pp. 1-17.

[22] V. K. JAIN AND H. M. SRIVASTAVA, *q-Series identities and reducibility of basic double hypergeometric functions*, Canad. J. Math., 38 (1986), pp. 215-231.

[23] B. NASSRALLAH, Ph.D. thesis, Carleton University, Ottawa, Ontario, 1982.
[24] M. RAHMAN AND A. VERMA, *Product and addition formulas for continuous q-ultraspherical polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 1461-1474.
[25] S. RAMANUJAN, Collected Papers, Cambridge, University Press, Cambridge, 1927, pp. 23-39.
[26] ———, *Notebook, Vol.* 2, Tata Inst. Fund. Res., Bombay, 1957.
[27] L. J. ROGERS, *Third memoir on the expansion of certain infinite products*, Proc. Lond. Math. Soc., 26 (1895), pp. 15-32.
[28] V.N. SINGH, *The basic analogues of identities of the Cayley–Orr type*, J. Lond. Math. Soc., 34 (1959), pp. 15-22.
[29] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1283-1292, November 1989

DUALITY FOR FAMILIES OF NATURAL VARIATIONAL PRINCIPLES IN NONLINEAR ELECTROSTATICS*

J. F. TOLAND† AND J. R. WILLIS†

Abstract. Standard duality theory embeds a given minimisation problem into a family of such problems and relates the solution of the original problem to a particular member of the dual family of problems. This theory is extended, in the specific context of the nonlinear electrostatic theory of a composite dielectric material, by considering a family of minimisation problems, parametrised by the imposed data, whose solutions define a convex function of this data. Here, the primary variable is the electric field. A corresponding family of dual problems is constructed, in which the primary variable is the electric displacement. The novelty of the formulation is that the duality is interpreted in terms of the space of the parameters that define the original family, so that the dual family generates a convex function of a dual set of parameters. It is demonstrated, under mild hypotheses on the electric properties of the nonlinear dielectric material, that these two convex functions are Legendre transforms of each other. As a by-product, the precise complementary duality principle between individual pairs of variational problems (as opposed to families of variational problems) is elucidated.

Key words. duality, Legendre transforms, electrostatics

AMS(MOS) subject classifications. 49A55, 49A52, 49A27, 78A30

1. Introduction. This paper is concerned with two families of variational problems in the nonlinear electrostatic theory of a composite dielectric material. In one the dependent variable is the electric potential, and in the other family it is the electric displacement field. Each family is parametrised by the imposed data, and as such each minimisation problem defines a convex function of the space of imposed data. Our main goal is to show that, under very generous hypotheses on the electric properties of the nonlinear dielectric material, these two convex functions are Legendre transforms of one another. As a by-product, the precise complementary duality principle between individual pairs of variational problems (as opposed to the families of variational problems) is elucidated. The main result is explained in detail in § 3.4, but it is appropriate to give a sketch of it here.

Let Ω denote a connected open bounded set containing a composite material with variable dielectric constant. The material may contain regions of conductor where the dielectric constant is infinite, and regions of insulator where the dielectric constant is zero. Let Γ_φ denote a prescribed nonplanar portion of the boundary Γ of Ω, and let H_p denote a space of functions on Ω vanishing in a weak sense on Γ_φ. Then the primary family of boundary value problems may be written loosely as

$$\min_{\varphi \in H_p} \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, \nabla\varphi(x) + \Lambda)\, dx = F(\Lambda). \tag{1.1}$$

Here $\mathbf{E}: \Omega \times \mathbf{R}^3 \to \bar{\mathbf{R}} = \mathbf{R} \cup \{+\infty\}$ denotes the electric energy density function (whose properties are described in detail in § 3.2), $\Lambda \in \mathbf{R}^3$, and the equilibrium potential $\Phi = \varphi + \Lambda.x$ is $\Lambda.x$ in a weak sense on Γ_φ. The significance of this formulation is as follows. First, in the particular case that $\Gamma_\varphi = \Gamma$, Λ is the mean value over Ω of the equilibrium electric field; this case has direct relevance to composite media and is highlighted in [6] and [7], for example. Other cases may also have physical significance. For instance, if Γ_φ consists of portions of two distinct parallel planes and Φ is prescribed

* Received by the editors June 15, 1988; accepted for publication November 30, 1988.
† School of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom.

to be constant on each, then Λ provides a measure of the potential difference between the planes, although it is not, in general, the mean value over Ω of the electric field. In every case, $F(\Lambda)$ is the mean energy function associated with Λ. An alternative formulation is to let $\mathbf{E}^*(x, \cdot)$ be the Legendre transform of $\mathbf{E}(x, \cdot)$ and to pose a variational problem for the electric displacement which can be written loosely as

$$\min_{\substack{\operatorname{Div} w^*=0,\\ (1/|\Omega|)\int_\Omega w^*(x)\,dx=\lambda^*,\\ w^*.n=0 \quad \text{on } \Gamma\setminus\Gamma_\varphi}} \frac{1}{|\Omega|}\int_\Omega \mathbf{E}^*(x, w^*(x))\,dx = f_*(\lambda^*). \tag{1.2}$$

A noteworthy feature of problem (1.2) is that no boundary data are imposed in the case $\Gamma = \Gamma_\varphi$; λ^* prescribes only the mean of w^* over Ω.

Our purpose is to say precisely that f_* and F are Legendre duals of one another and to establish (3.12) and (3.13) when the variational principles are correctly formulated in Sobolev spaces over a domain Ω with sufficient regularity on the boundary. It is important to recognise that this result is different from one that says that two variational problems (as opposed to two families of variational problems) are related by duality. As a consequence of the present considerations we find an answer to the following question. If, for given $\Lambda \in \mathbf{R}^3$, φ is a solution of (1.1), what is the mean electric displacement that must be imposed in problem (1.2) to obtain the same equilibrium? The solution is given in § 3.4.

This paper puts the work of Talbot and Willis [6] and Willis [7] in its application to electrostatics in a precise functional analytic context. These authors seem to have been the first to address the question of duality for families of variational problems parametrised by spaces of prescribed data. In fact, the duality theory is implicit in the work of Ekeland and Temam [2], although § 2 of this article seems to be the first place where the present implications of classical Fenchel duality theory have been explicitly recognised for a generally parametrised family of variational problems. Clearly, the work of § 3 is only one special case where this viewpoint is valuable. Extensions to certain others, as addressed, for example, by Talbot and Willis [6], Willis [7], and Ponte-Castañeda and Willis [4] would be straightforward. On the other hand, an extension to finite deformation nonlinear elasticity would be less so, in view of the need to relax convexity assumptions (see, e.g., Ball [1]). This latter problem is a subject of ongoing research.

2. The duality theory. Let V, V^*; W, W^*; and X, X^* be three pairs of topological vector spaces in separating duality, the duality pairing being denoted by $\langle , \rangle$, and let

$$\mathscr{F}\colon V \times W \times X \to \bar{\mathbf{R}} \quad \text{(the extended reals)}$$

be a proper, convex, lower semicontinuous function (proper means $\mathscr{F} \not\equiv -\infty$ anywhere, and $\mathscr{F} \not\equiv +\infty$). For fixed $\Lambda \in X$, let

$$F(\Lambda) = \inf_{v \in V} \mathscr{F}(v, 0, \Lambda). \tag{2.1}$$

Now define $\Phi_\Lambda\colon V \times W \times X \to \bar{\mathbf{R}}$ by

$$\Phi_\Lambda(v, w, \lambda) = \mathscr{F}(v, w, \Lambda + \lambda).$$

Then

$$F(\Lambda) = \inf_{u \in V} \Phi_\Lambda(u, 0, 0). \tag{2.2}$$

By standard Fenchel duality (see Ekeland and Temam [2])

$$F(\Lambda) \geqq \sup_{(w^*,\lambda^*)\in W^*\times X^*} -\Phi_\Lambda^*(0, w^*, \lambda^*), \tag{2.3}$$

whence

$$\begin{aligned} F(\Lambda) &\geqq \sup_{(w^*,\lambda^*)\in W^*\times X^*} \{\langle \lambda^*, \Lambda\rangle - \mathscr{F}^*(0, w^*, \lambda^*)\} \\ &= \sup_{\lambda^*\in X^*} \{\langle \lambda^*, \Lambda\rangle - \inf_{w^*\in W^*} \mathscr{F}^*(0, w^*, \lambda^*)\}. \end{aligned} \tag{2.4}$$

Here $\mathscr{F}^*$ denotes the Legendre transform of $\mathscr{F}$ (which is called the polar function of $\mathscr{F}$ by Ekeland and Temam [2]).

Now let

$$F_*(\lambda^*) = \inf_{w^*\in W^*} \mathscr{F}^*(0, w^*, \lambda^*).$$

Then (2.4) says that

$$F(\Lambda) \geqq \sup_{\lambda^*\in X^*} \{\langle \lambda^*, \Lambda\rangle - F_*(\lambda^*)\}.$$

In other words,

$$F(\Lambda) \geqq (F_*)^*(\Lambda), \qquad \Lambda \in X. \tag{2.5}$$

Taking the Legendre transform of both sides of (2.5) gives

$$F^*(\Lambda^*) \leqq (F_*)^{**}(\Lambda^*) \leqq F_*(\Lambda^*), \qquad \Lambda^* \in X^*. \tag{2.6}$$

We now apply standard theory to the question of equality in (2.5). We adopt the convention that equality in (2.5) means that both quantities are finite and equal for all $\Lambda \in X$. There follows the well-known criterion for equality (Ekeland and Temam [2, Chap. III, Prop. 2.1]).

PROPOSITION 2.1. *Equality in* (2.5) *holds if and only if* $h_\Lambda(0, 0)$ *is finite and* h_Λ *is lower semicontinuous at* $(0, 0)$, *where*

$$h_\Lambda(w, \lambda) = \inf_{v\in V} \mathscr{F}(v, w, \Lambda + \lambda). \tag{2.7}$$

It is interesting to ask whether equality in (2.6) follows from equality in (2.5). This will be so if and only if F_* is a proper convex lower semicontinuous function of Λ^*. The next proposition gives hypotheses when it is.

Let $\Lambda^* \in X^*$, $w \in W$, and

$$H_{\Lambda^*}(w) = \inf_{(v,\Lambda)\in V\times X} \{\mathscr{F}(v, w, \Lambda) - \langle \Lambda^*, \Lambda\rangle\}.$$

PROPOSITION 2.2. *The function* $F_*: X^* \to \bar{\mathbf{R}}$ *is a proper convex lower semicontinuous function if* $H_{\Lambda^*}(0)$ *is finite and* H_{Λ^*} *is lower semicontinuous at* $0 \in W$ *for each* $\Lambda^* \in X^*$.

Proof. Let Λ^* be fixed,

$$\Psi(v^*, w^*, \lambda^*) = \mathscr{F}^*(v^*, w^*, \Lambda^* + \lambda^*)$$

and note that (since $\mathscr{F}^{**} = \mathscr{F}$)

$$\Psi^*(v, w, \Lambda) = \mathscr{F}(v, w, \Lambda) - \langle \Lambda^*, \Lambda\rangle.$$

By hypothesis,

$$H_{\Lambda^*}(w) = \inf_{(v,\Lambda)\in V\times X} \Psi^*(v, w, \Lambda)$$

is finite and lower semicontinuous at $0 \in W$. Hence by [2, Chap. III, Prop. 2.1],

$$\hat{H}(v^*, \lambda^*) = \inf_{w^* \in W^*} \mathscr{F}^*(v^*, w^*, \Lambda^* + \lambda^*)$$

defines a function which is finite and lower semicontinuous at $(0,0) \in V^* \times X^*$. A fortiori, $\hat{H}(0, \cdot): X^* \to \bar{\mathbf{R}}$ is finite and lower semicontinuous at $0 \in X^*$. But $\hat{H}(0, \lambda^*) = F_*(\Lambda^* + \lambda^*)$. Hence F_* is finite and lower semicontinuous on X^*. □

Let us suppose that the hypotheses of both propositions hold. We then conclude that

$$F(\Lambda) = (F_*)^*(\Lambda), \qquad \Lambda \in X \tag{2.8}$$

and

$$F^*(\Lambda^*) = F_*(\Lambda^*), \qquad \Lambda^* \in X^*. \tag{2.9}$$

Moreover,

$$F(\Lambda) + F_*(\Lambda^*) \geqq \langle \Lambda, \Lambda^* \rangle, \tag{2.10}$$

and

$$F(\Lambda) + F_*(\Lambda^*) = \langle \Lambda, \Lambda^* \rangle \tag{2.11}$$

if and only if either one of the following two extremality conditions hold:

$$\Lambda^* \in \partial F(\Lambda) \quad \text{or} \quad \Lambda \in \partial F_*(\Lambda^*). \tag{2.12}$$

Remark. If F and F_* are defined and finite on a finite-dimensional space, then the subdifferentials are everywhere nonempty. In § 3.4 the significance of these observations is made clear in a particular example.

Now we are in a position to formulate the electrostatic boundary value problems sufficiently precisely so that the results indicated in § 1 can be inferred from the above interpretation of duality theory.

3. Mathematical formulation.

3.1. The composite. In this section we specify the domain occupied by the composite and its dielectric properties in a sufficiently precise way so that the duality theory of the preceding section can be invoked in appropriate spaces of admissible functions. What we have in mind is a body of composite material possibly containing interior regions of conducting material where the dielectric constant is infinite and regions of insulating material where the dielectric constant is zero.

Let $\Omega \subset \mathbf{R}^3$ be a bounded, open connected set which represents the region occupied by the composite, including insulating and conducting regions. Let Γ denote the Lipschitz boundary of Ω. Suppose that the closed subset F of Ω occupied by insulator is the union of a set of zero measure and the closure of an open set with Lipschitz continuous boundary. (It is possible to weaken this assumption on boundary regularity by supposing only minimal smoothness in the sense of Stein [5, Chap. VI, § 3].) The conducting region in Ω will be denoted by G, and it is a closed set which does not intersect F. Let $\Gamma_\varphi \subset \Gamma$ denote a subset of Γ of positive two-dimensional measure, which is not a subset of a plane. (Γ_φ may contain flat portions of the boundary Γ, provided it is not entirely coplanar.)

3.2. The electric energy density function. The electric displacement D may be written formally in terms of the electric field E as $D = \mathbf{E}'(x, E)$, where $'$ denotes differentiation with respect to E, and $x \in \Omega$. To pose the electrostatic boundary value

problem variationally there is no need for $\mathbf{E}$ to be differentiable, but the regularity of solutions of the variational problem is determined by the differentiability of $\mathbf{E}$. We assume, first of all, that $\mathbf{E}$, which is possibly infinite at points $x \in G$, is a Carathéodory function:

(i) For all $E \in \mathbf{R}^3$, $x \to \mathbf{E}(x, E)$ is measurable on Ω;

(ii) For almost all $x \in \Omega$, $E \to \mathbf{E}(x, E)$ is continuous on $\mathbf{R}^3$.

In addition we suppose

(iii) $E \to \mathbf{E}(x, E)$ is a proper, convex, extended real-valued function for almost all $x \in \Omega$;

(iv)

$$\left| \int_\Omega \mathbf{E}(x, u(x))\, dx \right| < \infty$$

for all smooth functions u such that $u(x) = 0, x \in G$;

(v)

$$\mathbf{E}(x, E) = 0, \quad x \in F, \quad E \in \mathbf{R}^3;$$

(vi) There exist $q > 1$, $b : \Omega \to (0, \infty)$, and $c \in L_1(\Omega)$ such that

$$\mathbf{E}(x, E) \geqq b(x)|E|^q - c(x), \qquad x \in \Omega \backslash F,$$

where $1/b \in L_\alpha(\Omega)$, for some $\alpha > 1/(q-1)$.

Conditions (i) and (ii) are more or less unavoidable in a modern treatment of the calculus of variations; (iii)-(vi) ensure that the variational problem is bounded below and is convex, and guarantees the lower semicontinuity needed to invoke the theory of § 2 in certain Sobolev spaces of pth power integrable functions where

$$p = \frac{\alpha q}{1+\alpha} > 1. \tag{3.1}$$

Condition (iv) allows Ω to contain a region G of conducting material. We simply put

$$\mathbf{E}(x, 0) = 0, \qquad x \in G,$$

$$\mathbf{E}(x, E) = +\infty, \quad x \in G, \quad E \neq 0.$$

The integrand in the variational problem then penalises the conducting region naturally and forces $E = 0$ in G. Condition (v) allows Ω to contain regions of insulator.

3.3. Function spaces. Let $\hat{C}_\infty$ denote the space of all smooth functions on Ω that are zero in a neighbourhood of Γ_φ, and let $\hat{H}_p$ denote the completion of $\hat{C}_\infty$ with respect to the norm

$$\|u\|_p^p = \int_\Omega \{|u(x)|^p + |\nabla u(x)|^p\}\, dx.$$

Throughout this section p is given by the formula (3.1). We note, immediately, that if $u \in \hat{C}_\infty$, then Poincaré's inequality [3, Thm. 3.6.4] yields

$$\int_\Omega |u(x)|^p\, dx \leqq (\text{const}) \int_\Omega |\nabla u(x)|^p\, dx.$$

Hence the norm $\|\cdot\|_p$ is equivalent to $|\cdot|_p$ where

$$|u|_p^p = \int_\Omega |\nabla u(x)|^p\, dx, \qquad u \in \hat{H}_p.$$

Moreover, it follows that the space

$$V_p = \{E = \nabla\varphi : \varphi \in \hat{H}_p\}$$

is a closed subspace of $(L_p(\Omega))^3$. Since Γ_φ is not a subset of a plane, we know that the only constant function in V_p is zero.

Now we are in a position to formulate the electrostatic boundary value problem in a precise way.

3.4. Boundary value problems. The family of primal problems parametrised by $\Lambda \in \mathbf{R}^3$ is

$$\mathscr{P}_\Lambda: \inf_{E \in V_p} \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, \Lambda + E(x))\, dx. \tag{3.2}$$

To put this in the context of § 2, we define our spaces of functions as follows:

$X = \{\Lambda : \Lambda \in \mathbf{R}^3\}$, the constant functions in $(L_p(\Omega))^3$;
$V = V_p$, a closed subspace of $(L_p(\Omega))^3$ with $V \cap X = \{0\}$;
$W =$ any topological complement of $X \oplus V$ in $(L_p(\Omega))^3$;
$X^* = (V \oplus W)^\perp \subset (L_{p'}(\Omega))^3$;
$V^* = (X \oplus W)^\perp \subset (L_{p'}(\Omega))^3$;
$W^* = (X \oplus V)^\perp \subset (L_{p'}(\Omega))^3$;

where $p^{-1} + p'^{-1} = 1$ and $\perp$, as usual, means the annihilator. Each is a normed linear space with respect to the norm inherited from $(L_p(\Omega))^3$ or $(L_{p'}(\Omega))^3$. Let $\langle u, v\rangle = (1/|\Omega|) \int_\Omega u(x) \cdot v(x)\, dx$, $(u, v) \in (L_p(\Omega))^3 \times (L_{p'}(\Omega))^3$.

LEMMA 3.1. *Each of* (X, X^*), (V, V^*), *and* (W, W^*) *is a pair of Banach spaces in separating duality (using the dual pairing of* $(L_p(\Omega))^3$ *and* $(L_{p'}(\Omega))^3$ *given above), and* X^* *has dimension three.*

Proof. Because of the symmetry in the definitions, it will suffice to show that the result holds for (W, W^*). If $w \in W$, then by the Hahn-Banach theorem there exists a bounded linear functional f on $(L_p(\Omega))^3$ such that $f(w) \neq 0$ and $f(X \oplus V) = 0$, since $X \oplus V$ is a closed subspace. Hence $f \in W^*$ and $f(w) \neq 0$. Now suppose $w^* \in W^*$. If $w^* \neq 0$, then there exists $u \in (L_p(\Omega))^3$ such that $w^*(u) \neq 0$. But $u = v + w + x \in V \oplus W \oplus X$, and so $w^*(w) \neq 0$ since $w^*(v + x) = 0$. Hence (W, W^*) are in separating duality.

Since X has dimension three, let $X = \operatorname{span}\{e_1, e_2, e_3\}$. By the Hahn-Banach theorem there exist three bounded linear functionals f_1, f_2, and f_3 on $(L_p(\Omega))^3$ such that $f_i(e_j) = \delta_{ij}$ and $f_i(V \oplus W) = 0$, $i = 1, 2, 3$. Hence these are three linearly independent elements of X^*. Now suppose that $f \in X^*$ and $x = \alpha e_1 + \beta e_2 + \gamma e_3 \in X$. Then $f(V \oplus W) = 0$, and $f(x) = f(\alpha e_1 + \beta e_2 + \gamma e_3) = f(e_1) f_1(x) + f(e_2) f_2(x) + f(e_3) f_3(x)$ and so $f \in \operatorname{span}\{f_1, f_2, f_3\}$. □

Let us suppose now that, with respect to the function spaces just specified, the functional

$$\mathscr{F}(v, w, \Lambda) = \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, v(x) + w(x) + \Lambda)\, dx$$

satisfies the hypotheses of the propositions of § 2. (The proof that this is so under the assumptions of §§ 3.1 and 3.2 is the substance of § 4.) To calculate the dual family of problems parametrised by $\Lambda^* \in X^*$, we need to calculate the Legendre transform of $\mathscr{F}$. Now $|\mathscr{F}(0, 0, 0)| < \infty$ by condition (iv), and so, by [2, Chap. IX, Prop. 2.1], we find

that

$$\begin{aligned}
&\mathscr{F}^*(0, w^*, \Lambda^*) \\
&\quad = \sup_{(v,w,\Lambda)\in V\times W\times X} \left\{ \langle \Lambda, \Lambda^* \rangle + \frac{1}{|\Omega|} \int_\Omega \{ w(x)\cdot w^*(x) - \mathbf{E}(x, v(x)+w(x)+\Lambda)\}\, dx \right\} \\
&\quad = \sup_{(v,w,\Lambda)\in V\times W\times X} \frac{1}{|\Omega|} \int_\Omega \{ (v+w+\Lambda)\cdot(\Lambda^*+w^*) - \mathbf{E}(x, v+w+\Lambda\}\, dx \\
&\quad = \sup_{u\in (L_p(\Omega))^3} \frac{1}{|\Omega|} \int_\Omega \{ u\cdot(\Lambda^*+w^*) - \mathbf{E}(x, u)\}\, dx \\
&\quad = \frac{1}{|\Omega|} \int \mathbf{E}^*(x, \Lambda^*(x)+w^*(x))\, dx,
\end{aligned}$$

where $\mathbf{E}^*(x, u^*) = \sup_{u\in\mathbf{R}^3} \{u\cdot u^* - \mathbf{E}(x, u)\}$ for all $u^*\in\mathbf{R}^3$. (Here $\cdot$ denotes the usual inner product in $\mathbf{R}^3$.)

Hence, in the light of § 3.3, the dual family is parametrised by the three-dimensional linear space X^* as follows:

$$\mathscr{P}^*_{\Lambda^*}\colon \inf_{w^*\in W^*} \frac{1}{|\Omega|} \int_\Omega \mathbf{E}^*(x, \Lambda^*(x)+w^*(x))\, dx, \tag{3.3}$$

and for each $\Lambda^*\in X^*$ we let $F_*(\Lambda^*)$ denote the right-hand side of (3.3). The physical significance of $P^*_{\Lambda^*}$ may be elucidated by the following observations.

If $x\in F$, then $\mathbf{E}(x, E) = 0$ for all E, and so

$$\mathbf{E}^*(x, 0) = 0, \quad \mathbf{E}^*(x, u^*) = +\infty, \quad u^*\neq 0.$$

As a consequence, any $w^*\in W^*$ where the integral in (3.3) is finite has the property that $\Lambda^*(x)+w^*(x) = 0$ almost everywhere on F, the region containing the insulator. If $\Lambda^*\in X^*$ is fixed, then (3.3) is finite only if, in a weak sense,

$$\operatorname{Div}(\Lambda^*+w^*) = 0 \quad \text{on } \Omega, \tag{3.4}$$

$$(\Lambda^*+w^*) = 0 \quad \text{on } F, \tag{3.5}$$

$$(\Lambda^*+w^*).n = 0 \quad \text{on } \Gamma\backslash\Gamma_\varphi, \tag{3.6}$$

and

$$\frac{1}{|\Omega|} \int_\Omega (\Lambda^*(x)+w^*(x))\, dx = \frac{1}{|\Omega|} \int_\Omega \Lambda^*(x)\, dx, \tag{3.7}$$

since, by definition of the spaces,

$$\int_\Omega (\Lambda^*(x)+w^*(x))\cdot\nabla\varphi(x)\, dx = 0, \quad \varphi\in\hat{H}_p \quad \text{and} \quad \int_\Omega w^*(x)\, dx = 0. \tag{3.8}$$

Because of Lemma 3.1, we know that the mapping

$$\Lambda^* \to \frac{1}{|\Omega|} \int_\Omega \Lambda^*(x)\, dx \tag{3.9}$$

is a linear bijection from X^* onto $\mathbf{R}^3$. Therefore, the family of dual principles can

be rewritten in terms of variational problems for electric displacements with mean $\lambda^* = 1/|\Omega| \int_\Omega \Lambda^*(x)\,dx$ as follows (see (1.2)):

$$\mathscr{P}_{\lambda^*}: \inf_{\substack{u^* \in (L_{p'}(\Omega))^3, \\ \operatorname{Div} u^* = 0, \\ (1/|\Omega|)\int u^*(x)\,dx = \lambda^*, \\ u^* \cdot n = 0 \quad \text{on } \Gamma \setminus \Gamma_\varphi}} \frac{1}{|\Omega|} \int_\Omega \mathbf{E}^*(x, u^*(x))\,dx. \tag{3.10}$$

Note that for $(\Lambda, \Lambda^*) \in X \times X^*$,

$$\langle \Lambda, \Lambda^* \rangle = \frac{1}{|\Omega|} \int_\Omega \Lambda \cdot \Lambda^*(x)\,dx = \Lambda \cdot \lambda^*, \tag{3.11}$$

and note that the primary family of problems can be written in a weak sense as

$$\mathscr{P}_\Lambda: \inf_{\substack{\varphi \in W^{1,p}(\Omega), \\ \varphi = \Lambda \cdot x \quad \text{on } \Gamma_\varphi}} \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, \nabla\varphi(x))\,dx.$$

(This means that φ is admissible if $(\varphi - \Lambda \cdot x) \in \hat{H}_p$.) The duality theory of § 2 can now be written out in the following weak sense:

$$F^*(\lambda^*) = f_*(\lambda^*) \quad \text{and} \quad (f_*)^*(\Lambda) = F(\Lambda), \tag{3.12}$$

i.e.,

$$\inf_{\substack{\varphi \in W^{1,p}(\Omega), \\ \varphi = \Lambda \cdot x \quad \text{on } \Gamma}} \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, \nabla\varphi(x))\,dx + \inf_{\substack{u^* \in (L_{p'}(\Omega))^3, \\ \operatorname{Div} u^* = 0, \\ (1/|\Omega|)\int u^*(x)\,dx = \lambda^*, \\ u^* \cdot n = 0 \quad \text{on } \Gamma \setminus \Gamma_\varphi}} \frac{1}{|\Omega|} \int_\Omega \mathbf{E}^*(x, u^*(x))\,dx \geqq \Lambda \cdot \lambda^*$$

and equality holds if and only if

$$\lambda^* \in \partial F(\Lambda) \quad \text{and} \quad \Lambda \in \partial f_*(\lambda^*), \tag{3.13}$$

where $f_*(\lambda^*)$ denotes the right-hand side of (3.9) and ∂F denotes the subdifferential of F.

To complete the discussion, we note that both F and f_* are defined on $\mathbf{R}^3$ and are convex. Hence at every point Λ, $\partial F(\Lambda) \neq \varnothing$, and at every point λ^*, $\partial f_*(\lambda^*) \neq \varnothing$. Moreover, under the hypotheses on $\mathbf{E}$, $F(\Lambda)$ and $f_*(\lambda^*)$ are both attained at points of $\hat{H}_p$ and W^*.

The physical significance of these observations is an answer to the following question. If $\Lambda \in \mathbf{R}^3$ and $\mathscr{P}_\Lambda$ has equilibrium potential given by $\Lambda \cdot x + \varphi$, $\varphi \in \hat{H}_p$, what mean electric displacement must be prescribed to obtain *the same equilibrium* in this nonlinear problem? The answer is obtained by solving the first equation in (3.13).

Note that, because of the growth condition (vi) on the electric energy density function, there always exists $\varphi \in H_p$ such that $F(\Lambda) = \mathscr{F}(\Lambda + \nabla\varphi)$.

4. Lower semicontinuity. This section contains proofs of the technical results needed to invoke the theory of § 2 for the nonlinear dielectric functions. Throughout we will suppose that Ω and $\mathbf{E}$ are as specified in §§ 3.1 and 3.2, and that $p = \alpha q/(1+\alpha) > 1$. Let $\mathscr{F}: (L_p(\Omega))^3 \to \bar{\mathbf{R}}$ be defined by

$$\mathscr{F}(u) = \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, u(x))\,dx, \qquad u \in (L_p(\Omega))^3. \tag{4.1}$$

Because of conditions (i)-(vi) in § 3, $\mathscr{F}$ is a proper convex function, and our first task is to show that it is lower semicontinuous with respect to the norm topology on $(L_p(\Omega))^3$. Suppose $u_n \to u$ in $(L_p(\Omega))^3$. Then, extracting a subsequence if necessary, we may suppose that $u_n \to u$ pointwise almost everywhere. Since $\mathbf{E}$ is bounded below on $\Omega\backslash F$ by the integrable function $-c$ and condition (v) holds, it is immediate from Fatou's lemma that

$$\mathscr{F}(u) \leqq \liminf \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, u_n(x))\, dx.$$

In other words, $\mathscr{F}$ is lower semicontinuous on $(L_p(\Omega))^3$. Because $\mathscr{F}$ is convex it is lower semicontinuous with respect to weak convergence as well. Hence we can infer (2.5) and (2.6) without making any growth assumptions (such as condition (vi)) on $\mathbf{E}$. But we need to verify the hypotheses of Propositions 2.1 and 2.2 to get equality in these inequalities.

Let $\Lambda \in X$ and let h_Λ be defined by (2.7). Then $h_\Lambda(0, 0) < \infty$ by condition (iv). Let Λ be fixed, and let $(w_n, \lambda_n) \in W \times X$ be such that $(w_n, \lambda_n) \to (0, 0)$, and let $v_n \in V$ be such that

$$(4.2) \quad \mathscr{F}(v_n + w_n + \lambda_n + \Lambda) = \frac{1}{|\Omega|} \int_\Omega \mathbf{E}(x, v_n + w_n + \lambda_n + \Lambda)\, dx \leqq h_\Lambda(w_n, \lambda_n) + 1/n.$$

Now the boundary of $(\Omega\backslash F)$ is Lipschitz, and so by Stein's extension theory [5, Chap. VI, § 3], there is no loss of generality in supposing that

$$\|v_n\|_{(L_p(\Omega))^3} \leqq (\text{const}) \|v_n\|_{(L_p(\Omega\backslash F))^3} \quad \text{since } \mathbf{E}(x, E) = 0, \quad x \in F.$$

If $h_\Lambda(w_n, \lambda_n) \to \infty$ as $n \to \infty$, then it is trivial that $h_\Lambda(0, 0) \leqq \liminf h_\Lambda(w_n, \lambda_n)$. So suppose $h_\Lambda(w_n, \lambda_n)$ is bounded, by M, say. Then by (4.2) and condition (vi), putting $v_n + w_n + \lambda_n + \Lambda = u_n$, we get

$$\int_{\Omega\backslash F} |u_n(x)|^p\, dx = \int_{\Omega\backslash F} |[b(x)^{p/q} |u_n(x)|^p](b(x))^{-p/q}|\, dx$$

(by Hölder's inequality)

$$\leqq \left\{ \int_{\Omega\backslash F} b(x) |u_n(x)|^q\, dx \right\}^{p/q} \left\{ \int_{\Omega\backslash F} b(x)^{-p/(q-p)}\, dx \right\}^{(q-p)/q}$$

$$\leqq \left\{ \int_\Omega \mathbf{E}(x, u_n(x))\, dx + \|c\|_{L_1(\Omega)} \right\}^{p/q} \left\{ \int_\Omega \left| \frac{1}{b(x)} \right|^\alpha \right\}^{(q-p)/q},$$

since $\mathbf{E}(x, E) = 0$, $x \in F$. Hence $\|u_n\|_{(L_p(\Omega\backslash F))^3}$ is bounded. Since $w_n + \lambda_n \to 0$ in $(L_p(\Omega))^3$, we conclude that (v_n) is bounded in $(L_p(\Omega\backslash F))^3$ and hence in $(L_p(\Omega))^3$. Since $p > 1$, (v_n) has a weakly convergent subsequence, $v_n \overset{w}{\to} v$, say, and $v \in V$ because V is weakly closed. Since $\mathscr{F}$ is weakly lower semicontinuous in $(L_p(\Omega))^3$ we find that

$$h_\Lambda(0, 0) \leqq \mathscr{F}(v + \Lambda) \leqq \liminf \mathscr{F}(v_n + w_n + \lambda_n + \Lambda)$$

$$\leqq \liminf h_\Lambda(w_n, \lambda_n),$$

which proves that the hypotheses of Proposition 2.1 hold.

Finally, to verify the hypotheses of Proposition 2.2, we note that the argument for H_{Λ^*} being finite and for H_{Λ^*} being lower semicontinuous is identical to the one that we have just given. □

REFERENCES

[1] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 65 (1977), pp. 193–281.

[2] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[3] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1966.

[4] P. PONTE-CASTAÑEDA AND J. R. WILLIS, *On the overall properties of nonlinearly viscous composites*, Proc. Royal Soc. London Ser. A, 416, 1988, pp. 217–244.

[5] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[6] D. R. S. TALBOT AND J. R. WILLIS, *Bounds and self-consistent estimates for the overall properties of nonlinear composites*, IMA J. Appl. Math., 39 (1987), pp. 215–240.

[7] J. R. WILLIS, *The structure of overall constitutive relations for a class of nonlinear composites*, IMA J. Appl. Math., to appear.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1293–1320, November 1989

002

ELECTROSTATIC PROBLEMS FOR TWO CONDUCTING SPHERES*

ANDREW H. VAN TUYL†

Abstract. Investigations are carried out for two spheres at given potentials V_1 and V_2, and for two spheres at potential zero in the presence of an outside unit point charge. Integral representations involving elliptic functions are obtained for the solutions of these problems, starting from series solutions in dipolar coordinates. These integral representations are used to obtain the asymptotic behavior of the charge density as the distance ε between the spheres tends to zero. Integral representations are found for the limiting charge density as $\varepsilon \to 0$, and convergent and asymptotic expansions for the limiting charge density are obtained. These results are used to investigate circles of zero charge density in the limit $\varepsilon \to 0$.

Key words. electrostatic problems, two spheres, potential, charge density

AMS(MOS) subject classification. 31B20

1. Introduction. The potential outside two charged conducting spheres was first given by Poisson [17] in 1811. Further investigations were carried out by Plana [16] in 1845, Kirchhoff [10] in 1861, and Carl Neumann [13] in 1863. Dipolar coordinates were used in [13], and the Green function for the exterior of two spheres was given for the first time. Later work concerning two charged spheres includes investigations by Jeffery [5], who developed the use of dipolar coordinates further, and Russell [18]–[21]. A detailed treatment of electrostatic problems for two spheres has been given by Kottler in [11]. More recently, electrostatic problems for two spheres have been of interest in connection with the conductivity of granular materials (Keller [9], Batchelor and O'Brien [2], and Jeffrey [6], for example). In [7], Jeffrey has used an approach for nearly touching spheres based on the method of matched asymptotic expansions.

The present paper is concerned with investigations of the solutions of the electrostatic problems for two spheres defined by the following conditions: (1) The spheres are at given potentials V_1 and V_2; (2) The spheres are at potential zero in the presence of an outside point charge. These will be referred to as the first and second problems, respectively. The present investigation starts from the solutions of the first and second problems in dipolar coordinates in the form given by Ernst Neumann [14]. The solutions of these problems, including the potentials, charge densities, and total charges, are first expressed in terms of definite integrals that involve elliptic functions. From these integrals, definite integral representations are found for the limits of the charge densities in the first and second problems as the distance ε between the spheres goes to zero. Convergent and asymptotic expansions are obtained for the charge density at the inner axial points of the spheres as ε goes to zero, and for the limiting charge density.

Kirchhoff [10] found that the charge density at the inner axial points of two spheres with potentials and radii equal to unity is asymptotically proportional to $\varepsilon^{-3/2} \exp(-2^{-1}\pi^2\varepsilon^{-1/2})$ as ε tends to zero, correcting the orders of magnitude ε^2 and ε^3 given by Poisson [17] and Plana [16], respectively. In § 7 of the present paper, Kirchhoff's result is obtained from the integral representation for the charge density

* Received by the editors September 29, 1987; accepted for publication (in revised form) November 18, 1988. These results have been obtained at various times during the past 40 years, starting from portions of the author's unpublished Ph.D. thesis. Most of the results of §§ 4–6 were obtained at Indiana University in 1953 under Army Ordnance Contract DA-33-008 ord-454.

† Applied Mathematics Branch, Code R44, Naval Surface Warfare Center, Silver Spring, Maryland 20903-5000.

in § 6. In § 8, the corresponding result is found for the charge density in the second problem. It is found that the behavior with respect to ε remains the same as in the first problem, but with a coefficient which depends on the position of the point charge.

In § 10, the limiting charge densities in the first and second problems as $\varepsilon \to 0$ are expressed in terms of definite integrals. In §§ 11–14, these integrals are used to obtain various expansions, both convergent and asymptotic, for the limiting charge density. One of these expansions is a power series expansion in the neighborhood of the outer axial point. In § 15, this expansion is used to find the ratio $V_2/V_1 > 1$ such that, in the limit $\varepsilon \to 0$, all the charge on sphere 1 is negative while V_1 is positive.

2. Dipolar coordinates. Dipolar coordinates η, θ, and ϕ are defined by the equations

$$x + i\rho = ia \cot \tfrac{1}{2}(\theta + i\eta), \tag{2.1}$$

$$y = \rho \cos \phi, \qquad z = \rho \sin \phi, \tag{2.2}$$

with $a > 0$, $\rho > 0$. Let the points $(a, 0, 0)$ and $(-a, 0, 0)$ in Cartesian coordinates be denoted by A_1 and A_2, respectively. The coordinate surface $\eta =$ constant is a sphere with A_1 and A_2 as inverse points, and $\theta =$ constant is a spindle with A_1 and A_2 as vertices.

When $\eta_1 > 0$, the sphere $\eta = \eta_1$ has radius a csch η_1 and center at $x = a \coth \eta_1$, $y = z = 0$. Let P_1 be an arbitrary point on the sphere $\eta = \eta_1 > 0$, let its center be denoted by O_1, and let the angle between O_1P_1 and the negative x-axis be denoted by ω_1. Denoting the coordinates of P_1 by the subscript 1, we have

$$x_1 = a(\coth \eta_1 - \operatorname{csch} \eta_1 \cos \omega_1), \qquad \rho_1 = a \operatorname{csch} \eta_1 \sin \omega_1. \tag{2.3}$$

From (2.1) and (2.2), it follows that

$$\sin \theta_1 = \frac{\sinh \eta_1 \sin \omega_1}{\cosh \eta_1 - \cos \omega_1}, \tag{2.4}$$

with $\theta_1 = \pi$ when $\omega_1 = 0$ and $\theta_1 = 0$ when $\omega_1 = \pi$.

Given the two spheres $\eta = \eta_1$ and $\eta = \eta_2$, $\eta_1 > 0 > \eta_2$, the former contains A_1 in its interior and the latter contains A_2. Let their radii be r_1 and r_2, respectively, and let the distance between their centers be c. Since A_1 and A_2 are inverse with respect to both spheres, we have

$$\sqrt{r_1^2 + a^2} + \sqrt{r_2^2 + a^2} = c, \tag{2.5}$$

which yields

$$a = \sqrt{\left(\frac{c^2 - r_2^2 + r_1^2}{2c}\right)^2 - r_1^2} = \sqrt{\left(\frac{c^2 - r_1^2 + r_2^2}{2c}\right)^2 - r_2^2}. \tag{2.6}$$

Substituting $c = r_1 + r_2 + \varepsilon$, $\varepsilon > 0$, we find from either expression for a in (2.6) that

$$a = \frac{\sqrt{\varepsilon(\varepsilon + 2r_1)(\varepsilon + 2r_2)(\varepsilon + 2r_1 + 2r_2)}}{2(\varepsilon + r_1 + r_2)}. \tag{2.7}$$

In the special case $r_1 = r_2 = r$, (2.7) simplifies to

$$a = \tfrac{1}{2}\sqrt{\varepsilon(\varepsilon + 4r)}. \tag{2.8}$$

We also have

$$\eta_1 = \sinh^{-1} \frac{a}{r_1}, \qquad \eta_2 = -\sinh^{-1} \frac{a}{r_2}. \tag{2.9}$$

We can calculate a, η_1, and η_2 by use of (2.7)-(2.9) when r_1, r_2, and ε are known. When $\varepsilon \to 0$, we see that

$$a = \sqrt{\frac{2r_1r_2\varepsilon}{r_1+r_2}}[1+O(\varepsilon)], \tag{2.10}$$

$$\eta_1 = \sqrt{\frac{2r_2\varepsilon}{r_1(r_1+r_2)}}[1+O(\varepsilon)], \tag{2.11}$$

and

$$\eta_1 - \eta_2 = \sqrt{\frac{2(r_1+r_2)\varepsilon}{r_1r_2}}[1+O(\varepsilon)]. \tag{2.12}$$

From (2.4), (2.10), and (2.11), we find that

$$\theta_1 = \left(\frac{a}{r_1}\cot\frac{\omega_1}{2}\right)[1+O(\varepsilon)] \tag{2.13}$$

as $\varepsilon \to 0$. Finally, at a given point in space, we find from (2.1) that

$$\eta = \frac{2ax}{x^2+\rho^2}[1+O(\varepsilon)] \tag{2.14}$$

and

$$\theta = \frac{2a\rho}{x^2+\rho^2}[1+O(\varepsilon)] \tag{2.15}$$

as $\varepsilon \to 0$. Equation (2.7) does not appear to be in the literature.

3. Solutions of the electrostatic problems. In the coordinate system of § 2, let the given spheres be the coordinate surfaces $\eta = \eta_1$ and $\eta = \eta_2$, respectively, $\eta_1 > 0 > \eta_2$. In the first problem, we assume that the spheres are charged to constant potentials V_1 and V_2, respectively. In the second problem, the spheres are at zero potential in the presence of a unit point charge at the point $(\eta_0, \theta_0, \phi_0)$, where $\eta_1 > \eta_0 > \eta_2$. We will start with the solutions of these problems as given by Ernst Neumann in [14].

It is convenient to express the solutions of the first and second problems in terms of the function

$$I(\zeta, \theta) = \frac{1}{2}\sum_{n=0}^{\infty}\frac{\exp(N\zeta)}{\sinh N\delta}P_n(\cos\theta), \tag{3.1}$$

where $N = n+\frac{1}{2}$, $0 \leqq \theta \leqq \pi$, $P_n(\cos\theta)$ is the Legendre polynomial of degree n, $\delta > 0$, and $\operatorname{Re}\zeta < \delta$. We see that $I(\zeta, \theta)$ is an analytic function of ζ for $\operatorname{Re}\zeta < \delta$, and that (3.1) and

$$\frac{\partial I(\zeta, \theta)}{\partial\zeta} = \frac{1}{2}\sum_{n=0}^{\infty}\frac{N\exp(N\zeta)}{\sinh N\delta}P_n(\cos\theta) \tag{3.2}$$

are continuous with respect to θ at $\theta = 0$ and π when $\operatorname{Re}\zeta < \delta$. The latter follows immediately from uniform convergence with respect to θ in the interval $0 \leqq \theta \leqq \pi$, since we then have $|P_n(\cos\theta)| \leqq 1$.

Let the potential outside the spheres in the first problem be written in the form

$$V = \left(\frac{V_1+V_2}{2}\right)V^{(1)} + \left(\frac{V_1-V_2}{2}\right)V^{(2)}. \tag{3.3}$$

Then from [14] and (3.1), we have

$$V^{(1)} = \psi^{1/2}[I(\zeta_1, \theta) + I(-\zeta_1, \theta) - I(\zeta_2, \theta) - I(-\zeta_2-2\delta, \theta)], \tag{3.4}$$

and

$$V^{(2)} = \psi^{1/2}[I(\zeta_1, \theta) - I(-\zeta_1, \theta) + I(\zeta_2, \theta) - I(-\zeta_2 - 2\delta, \theta)], \tag{3.5}$$

where $\psi = 2(\cosh \eta - \cos \theta)$, $\delta = \eta_1 - \eta_2$, $\zeta_1 = \eta - \eta_1 - \eta_2$, and $\zeta_2 = \eta - \delta$. For the charge density on sphere 1, we have

$$D_1 = \left(\frac{V_1 + V_2}{2}\right) D_{11} + \left(\frac{V_1 - V_2}{2}\right) D_{12}, \tag{3.6}$$

where

$$D_{11} = \frac{\psi_1^{3/2}}{4\pi a} \frac{\partial}{\partial \zeta} [I(\zeta, \theta_1) + I(-\zeta, \theta_1)]|_{\zeta = -\eta_2} \tag{3.7}$$

and

$$D_{12} = \frac{\psi_1^{3/2}}{4\pi a} \frac{\partial}{\partial \zeta} [I(\zeta, \theta_1) - I(-\zeta, \theta_1)]|_{\zeta = -\eta_2}, \tag{3.8}$$

with $\psi_1 = 2(\cosh \eta_1 - \cos \theta_1)$. Finally, the coefficients of capacity and induction are given by

$$C_{11} = 2a \lim_{\theta \to 0} I(-\eta_1 - \eta_2, \theta), \tag{3.9}$$

$$C_{12} = -2a \lim_{\theta \to 0} I(-\delta, \theta), \tag{3.10}$$

and

$$C_{22} = 2a \lim_{\theta \to 0} I(\eta_1 + \eta_2, \theta). \tag{3.11}$$

The charge density on sphere 2 is obtained from (3.6)-(3.8) by interchanging subscripts 1 and 2 and reversing the signs of η_1 and η_2.

The potential satisfying the second problem is the Green function for the exterior of the spheres with respect to the point $(\eta_0, \theta_0, \phi_0)$. Denoting this potential by $G = 1/r + H$, we have

$$\begin{aligned} H = -\frac{\psi_0^{1/2}\psi^{1/2}}{2a} &[I(\zeta_1 + \eta_0, \gamma) + I(-\zeta_1 - \eta_0, \gamma) \\ &- I(\zeta_2 - \eta_0, \gamma) - I(-\zeta_2 + \eta_0 - 2\delta, \gamma)], \end{aligned} \tag{3.12}$$

where $\psi_0 = 2(\cosh \eta_0 - \cos \theta_0)$ and $\cos \gamma = \cos \theta \cos \theta_0 + \sin \theta \sin \theta_0 \cos (\phi - \phi_0)$. Finally, the charge density and total charge on sphere 1 are given by

$$D_1^* = -\frac{\psi_0^{1/2}\psi_1^{3/2}}{8\pi a^2} \frac{\partial}{\partial \zeta} [I(\zeta, \gamma_1) + I(-\zeta, \gamma_1)]|_{\zeta = \eta_0 - \eta_2} \tag{3.13}$$

and

$$Q_1^* = -\psi_0^{1/2}[I(\eta_0 - \eta_1 - \eta_2, \theta_0) - I(-\eta_0 - \delta, \theta_0)], \tag{3.14}$$

respectively. As before, we obtain D_2^* and Q_2^* by interchanging the subscripts 1 and 2 and reversing the signs of η_0, η_1, and η_2. We note that $-D_1^*$ and $-D_2^*$ form the kernel for the solution of the Dirichlet problem for the exterior of the two spheres.

4. Analytic continuation of $I(\zeta, \theta)$. When $0 < \theta < \pi$ and $\operatorname{Re} \zeta < \delta$, $\delta > 0$, we find that

$$\begin{aligned} I(\zeta, \theta) - I(\zeta - 2\delta, \theta) &= \sum_{n=0}^{\infty} \exp N(\zeta - \delta) P_n(\cos \theta) \\ &= [2 \cosh (\zeta - \delta) - 2 \cos \theta]^{-1/2}. \end{aligned} \tag{4.1}$$

We see that the right side of (4.1) is positive when ζ is real, and that it is single valued when the ζ-plane is cut along the line $\operatorname{Re}\zeta=\delta$ from $\delta+(2\pi n+\theta)i$ to $\delta+[(2n+2)\pi-\theta]i$ and from $\delta-(2\pi n+\theta)i$ to $\delta-[(2n+2)\pi-\theta]i$, $n\geqq 0$. Since $I(\zeta-2\delta,\theta)$ is analytic for $\operatorname{Re}\zeta<3\delta$, it follows that it is the analytic continuation of $I(\zeta,\theta)-[2\cosh(\zeta-\delta)-2\cos\theta]^{-1/2}$ in the region $\delta<\operatorname{Re}\zeta<3\delta$. Repeated application of (4.1) leads to the result

$$I(\zeta,\theta)=\sum_{r=1}^{m}\{2\cosh[\zeta-(2r-1)\delta]-2\cos\theta\}^{-1/2}+I(\zeta-2m\delta,\theta) \tag{4.2}$$

with $m\geqq 1$, valid for $\operatorname{Re}\zeta<(2m+1)\delta$ when the ζ-plane is cut from $(2r-1)\delta+(2n\pi+\theta)i$ to $(2r-1)\delta+[(2n+2)\pi-\theta)]i$ and from $(2r-1)\delta-(2n\pi+\theta)i$ to $(2r-1)\delta-[(2n+2)\pi-\theta]i$, $n\geqq 0$, $1\leqq r\leqq m$.

We have

$$|I(\zeta-2m\delta,\theta)|<\frac{|\exp(\zeta/2)|\exp(-m\delta)}{\sinh(\delta/2)} \tag{4.3}$$

for sufficiently large m. Hence, the expansion

$$I(\zeta,\theta)=\sum_{r=1}^{\infty}\{2\cosh[\zeta-(2r-1)\delta]-2\cos\theta\}^{-1/2} \tag{4.4}$$

holds for all ζ in the cut plane. Substitution of (4.4) in (3.3)-(3.5) gives the solution of the first problem as obtained by the method of images.

5. Integral representations for $I(\zeta,\theta)$. As before, let $\delta>0$ and $0<\theta<\pi$. When $-\delta<\operatorname{Re}\zeta<\delta$, we have

$$I(\zeta,\theta)+I(-\zeta,\theta)=\sum_{n=0}^{\infty}\frac{\cosh N\zeta}{\sinh N\delta}P_n(\cos\theta) \tag{5.1}$$

and

$$I(\zeta,\theta)-I(-\zeta,\theta)=\sum_{n=0}^{\infty}\frac{\sinh N\zeta}{\sinh N\delta}P_n(\cos\theta). \tag{5.2}$$

Substituting

$$P_n(\cos\theta)=\frac{2}{\pi}\int_{\theta}^{\pi}\frac{\sin Nt\,dt}{(2\cos\theta-2\cos t)^{1/2}} \tag{5.3}$$

[23, p. 315], in (5.1) and interchanging summation and integration, we obtain

$$I(\zeta,\theta)+I(-\zeta,\theta)=\frac{2}{\pi}\int_{\theta}^{\pi}(2\cos\theta-2\cos t)^{-1/2}\sum_{n=0}^{\infty}\frac{\cosh N\zeta\sin Nt}{\sinh N\delta}dt \tag{5.4}$$

when $-\delta<\operatorname{Re}\zeta<\delta$. The interchange of summation and integration is valid by [3, p. 495], since the series in (5.4) is uniformly convergent with respect to t for $\theta\leqq t\leqq\pi$, while the integral of $(2\cos\theta-2\cos t)^{-1/2}$ between θ and π is convergent. Similarly, from

$$P_n(\cos\theta)=\frac{2}{\pi}\int_{0}^{\theta}\frac{\cos Nt\,dt}{(2\cos t-2\cos\theta)^{1/2}} \tag{5.5}$$

[23, p. 315], and (5.2), we have

$$I(\zeta,\theta)-I(-\zeta,\theta)=\frac{2}{\pi}\int_{0}^{\theta}(2\cos t-2\cos\theta)^{-1/2}\sum_{n=0}^{\infty}\frac{\sinh N\zeta\cos Nt}{\sinh N\delta}dt \tag{5.6}$$

when $-\delta < \operatorname{Re}\zeta < \delta$. Finally, let k be defined by the equation

$$\delta = \pi K'/K, \tag{5.7}$$

where K is the complete elliptic integral of the first kind with modulus k and $K' = K(k')$, $k'^2 = 1 - k^2$. Then from (5.4), (5.6), and [23, p. 511], we obtain the integral representations

$$I(\zeta, \theta) + I(-\zeta, \theta) = \frac{Kk}{\pi^2}\int_\theta^\pi (2\cos\theta - 2\cos t)^{-1/2} \times \left[\operatorname{sn}\frac{K}{\pi}(t+i\zeta) + \operatorname{sn}\frac{K}{\pi}(t-i\zeta)\right] dt \tag{5.8}$$

and

$$I(\zeta, \theta) - I(-\zeta, \theta) = -i\frac{Kk}{\pi^2}\int_0^\theta (2\cos t - 2\cos\theta)^{-1/2} \times \left[\operatorname{sn}\frac{K}{\pi}(t+i\zeta) - \operatorname{sn}\frac{K}{\pi}(t-i\zeta)\right] dt \tag{5.9}$$

for $-\delta < \operatorname{Re}\zeta < \delta$, $0 < \theta < \pi$.

From (4.4), it follows that

$$I(\zeta - \delta, \theta) \pm I(-\zeta - \delta, \theta) = I(\zeta - \delta, \theta) \pm I(-\zeta + \delta, \theta) \mp (2\cosh\zeta - 2\cos\theta)^{-1/2} \tag{5.10}$$

for all ζ in the cut ζ-plane. When $\operatorname{Re}\zeta > 0$, we can verify that

$$\int_\theta^\pi (2\cos\theta - 2\cos t)^{-1/2}\left[\csc\frac{1}{2}(t+i\zeta) + \csc\frac{1}{2}(t-i\zeta)\right] dt = 2\pi(2\cosh\zeta - 2\cos\theta)^{-1/2} \tag{5.11}$$

and

$$\int_0^\theta (2\cos t - 2\cos\theta)^{-1/2}\left[\csc\frac{1}{2}(t+i\zeta) - \csc\frac{1}{2}(t-i\zeta)\right] dt = -2\pi i(2\cosh\zeta - 2\cos\theta)^{-1/2}. \tag{5.12}$$

The former can be shown by substituting $\cos(t/2) = \cos(\theta/2)\sin\phi$, and the latter, by substituting $\sin(t/2) = \sin(\theta/2)\sin\phi$. From (5.8)-(5.12), and using the addition theorem for sn u, we obtain

$$I(\zeta - \delta, \theta) + I(-\zeta - \delta, \theta) = \frac{K}{\pi^2}\int_\theta^\pi (2\cos\theta - 2\cos t)^{-1/2} \cdot \left\{\operatorname{ns}\frac{K}{\pi}(t+i\zeta) + \operatorname{ns}\frac{K}{\pi}(t-i\zeta) - \frac{\pi}{2K}\left[\csc\frac{1}{2}(t+i\zeta) + \csc\frac{1}{2}(t-i\zeta)\right]\right\} dt \tag{5.13}$$

and

$$I(\zeta - \delta, \theta) - I(-\zeta - \delta, \theta) = -\frac{iK}{\pi^2}\int_0^\theta (2\cos t - 2\cos\theta)^{-1/2} \cdot \left\{\operatorname{ns}\frac{K}{\pi}(t+i\zeta) - \operatorname{ns}\frac{K}{\pi}(t-i\zeta) - \frac{\pi}{2K}\left[\csc\frac{1}{2}(t+i\zeta) - \csc\frac{1}{2}(t-i\zeta)\right]\right\} dt \tag{5.14}$$

when $0 < \operatorname{Re}\zeta < 2\delta$. However, both sides of (5.13) and (5.14) are seen to be analytic for $-2\delta < \operatorname{Re}\zeta < 2\delta$, and hence, it follows by analytic continuation that they are valid

in the larger region. The preceding treatment can be extended to give integral representations which are valid in still wider strips of the ζ-plane.

Substituting $\cos(t/2)=\cos(\theta/2)\sin\phi$ in (5.8) and (5.13), and $\sin(t/2)=\sin(\theta/2)\sin\phi$ in (5.9) and (5.14), we find that

$$\lim_{\theta\to\pi}[I(\zeta,\theta)+I(-\zeta,\theta)]=\frac{Kk}{\pi}\operatorname{cd}\frac{iK\zeta}{\pi}, \tag{5.15}$$

$$\lim_{\theta\to 0}[I(\zeta,\theta)-I(-\zeta,\theta)]=-\frac{iKk}{\pi}\operatorname{sn}\frac{iK\zeta}{\pi}, \tag{5.16}$$

and

$$\lim_{\theta\to\pi}[I(\zeta-\delta,\theta)+I(-\zeta-\delta,\theta)]=\frac{K}{\pi}\left(\operatorname{dc}\frac{iK\zeta}{\pi}-\frac{\pi}{2K}\sec\frac{i\zeta}{2}\right), \tag{5.17}$$

$$\lim_{\theta\to 0}[I(\zeta-\delta,\theta)-I(-\zeta-\delta,\theta)]=-\frac{iK}{\pi}\left(\operatorname{ns}\frac{iK\zeta}{\pi}-\frac{\pi}{2K}\csc\frac{i\zeta}{2}\right). \tag{5.18}$$

We see that (5.8) and (5.13) remain valid for $\theta=0$, and that (5.9) and (5.14) remain valid for $\theta=\pi$. In particular, from (5.13), we have

$$\lim_{\theta\to 0}I(-\delta)=\frac{K}{2\pi^2}\int_0^{\pi}\left(\operatorname{ns}\frac{Kt}{\pi}-\frac{\pi}{2K}\csc\frac{t}{2}\right)\csc\frac{t}{2}\,dt. \tag{5.19}$$

6. Integral representations for the solutions of the electrostatic problems. We have $-\delta<\eta-\eta_1-\eta_2<\delta$ and $-2\delta<\eta<2\delta$ when $2\eta_2<\eta<2\eta_1$, where $\delta=\eta_1-\eta_2$. Then from (3.4), (3.5), and the preceding section, and using the addition theorem for sn u, we obtain

$$V^{(1)}=\frac{2K}{\pi^2}\psi^{1/2}\int_{\theta}^{\pi}(2\cos\theta-2\cos t)^{-1/2}\operatorname{Re}\left\{\operatorname{ns}\frac{K}{\pi}[(t+i(\eta-2\eta_1)]\right.$$
$$\left.-\operatorname{ns}\frac{K}{\pi}(t+i\eta)+\frac{\pi}{2K}\csc\frac{1}{2}(t+i\eta)\right\}dt \tag{6.1}$$

and

$$V^{(2)}=\frac{2K}{\pi^2}\psi^{1/2}\int_{0}^{\theta}(2\cos t-2\cos\theta)^{-1/2}\operatorname{Im}\left\{\operatorname{ns}\frac{K}{\pi}[t+i(\eta-2\eta_1)]\right.$$
$$\left.+\operatorname{ns}\frac{K}{\pi}(t+i\eta)-\frac{\pi}{2K}\csc\frac{1}{2}(t+i\eta)\right\}dt \tag{6.2}$$

when $2\eta_2<\eta<2\eta_1$. We see that the region exterior to the spheres, defined by $\eta_2<\eta<\eta_1$, is included in this inequality. It is easily verified that differentiation of the integrals in the preceding section with respect to ζ can be carried out under the integral sign. From (3.7), (3.8), (5.7), and (5.8), we therefore have

$$D_{11}=\frac{K^2k}{2\pi^4a}\psi_1^{3/2}\int_{\theta_1}^{\pi}(2\cos\theta_1-2\cos t)^{-1/2}\operatorname{Im}\operatorname{cn}\frac{K}{\pi}(t+i\eta_2)\operatorname{dn}\frac{K}{\pi}(t+i\eta_2)\,dt \tag{6.3}$$

and

$$D_{12}=\frac{K^2k}{2\pi^4a}\psi_1^{3/2}\int_{0}^{\theta_1}(2\cos t-2\cos\theta_1)^{-1/2}\operatorname{Re}\operatorname{cn}\frac{K}{\pi}(t+i\eta_2)\operatorname{dn}\frac{K}{\pi}(t+i\eta_2)\,dt. \tag{6.4}$$

From (3.9)-(3.11), and the results of the preceding section, noting that $-\delta < \eta_1 + \eta_2 < \delta$, we find that

$$C_{11} = \frac{aKk}{\pi^2}\int_0^{\pi} \operatorname{Re}\operatorname{sn}\frac{K}{\pi}[t + i(\eta_1+\eta_2)]\csc\frac{t}{2}\,dt + \frac{iaKk}{\pi}\operatorname{sn}\frac{iK}{\pi}(\eta_1+\eta_2), \tag{6.5}$$

$$C_{12} = -\frac{aK}{\pi^2}\int_0^{\pi}\left(\operatorname{ns}\frac{Kt}{\pi} - \frac{\pi}{2K}\csc\frac{t}{2}\right)\csc\frac{t}{2}\,dt, \tag{6.6}$$

and

$$C_{22} = C_{11} - \frac{2iaKk}{\pi}\operatorname{sn}\frac{iK}{\pi}(\eta_1+\eta_2). \tag{6.7}$$

Similarly, in the second problem, we have $-\delta < \eta - \eta_1 - \eta_2 + \eta_0 < \delta$ and $-2\delta < \eta - \eta_0 < 2\delta$ when $\eta_2 < \eta < \eta_1$ and $\eta_2 < \eta_0 < \eta_1$. Then as before, we have

$$\begin{aligned} H = -\frac{K}{\pi^2}\psi_0^{1/2}\int_{\gamma}^{\pi}(2\cos\gamma - 2\cos t)^{-1/2}\operatorname{Re}\Big\{&\operatorname{ns}\frac{K}{\pi}[t + i(\eta-\eta_1-\eta_2+\eta_0)] \\ &- \operatorname{ns}\frac{K}{\pi}[t+i(\eta-\eta_0)] + \frac{\pi}{2K}\csc\frac{1}{2}[t+i(\eta-\eta_0)]\Big\}\,dt, \end{aligned} \tag{6.8}$$

$$\begin{aligned} D_1^* = -\frac{K^2k}{4\pi^4a^2}\psi_0^{1/2}\psi_1^{3/2} &\cdot \int_{\gamma_1}^{\pi}(2\cos\gamma_1 - 2\cos t)^{-1/2}\operatorname{Im}\operatorname{cn}\frac{K}{\pi}[t+i(\eta_2-\eta_0)]\operatorname{dn}\frac{K}{\pi}[t+i(\eta_2-\eta_0)]\,dt, \end{aligned} \tag{6.9}$$

and

$$\begin{aligned} Q_i^* = -\frac{K}{\pi^2}\psi_0^{1/2}\Bigg\{&\int_{\theta_0}^{\pi}(2\cos\theta_0 - 2\cos t)^{-1/2} \\ &\cdot \operatorname{Re}\left\{\operatorname{ns}\frac{K}{\pi}[t+i(\eta_0-2\eta_1)] - \operatorname{ns}\frac{K}{\pi}(t+i\eta_0) + \frac{\pi}{2K}\csc\frac{1}{2}(t+i\eta_0)\right\}dt \\ &+ \int_0^{\theta_0}(2\cos t - 2\cos\theta_0)^{-1/2} \\ &\cdot \operatorname{Im}\left\{\operatorname{ns}\frac{K}{\pi}[t+i(\eta_0-2\eta_1)] + \operatorname{ns}\frac{K}{\pi}(t+i\eta_0) - \frac{\pi}{2K}\csc\frac{1}{2}(t+i\eta_0)\right\}dt\Bigg\} \end{aligned} \tag{6.10}$$

when $\eta_2 < \eta < \eta_1$ and $\eta_2 < \eta_0 < \eta_1$.

7. Behavior of D_{11} at the inner axial point as $\varepsilon \to 0$. Passing to the limit $\theta_1 = \pi$ in (6.3) by use of the substitution $\cos(t/2) = \cos(\theta_1/2)\sin\phi$, substituting $\eta_2 = \eta_1 - \pi K'/K$, and using the addition theorems for cn u and dn u, we obtain

$$D_{11}|_{\theta_1=\pi} = -i\frac{K^2kk'^2}{4\pi^3a}\psi_1^{3/2}\frac{\operatorname{sn}\dfrac{iK\eta_1}{\pi}}{\operatorname{cn}^2\dfrac{iK\eta_1}{\pi}}. \tag{7.1}$$

This expression was given in an equivalent form by Kirchhoff [10, p. 99], in terms of a modulus related to the present one by Landen's transformation. To obtain a series

for $D_{11}|_{\theta_1=\pi}$ which converges rapidly for small ε, we first find by use of Jacobi's imaginary transformation [23, p. 506] that

$$D_{11}|_{\theta_1=\pi}=\frac{K^2k'^2}{4\pi^3a}\psi_1^{3/2}\operatorname{sn}\left(\frac{K\eta_1}{\pi},k'\right)\operatorname{cn}\left(\frac{K\eta_1}{\pi},k'\right). \tag{7.2}$$

Noting that $k'^2\operatorname{sn}(u,k')\operatorname{cn}(u,k')=-(d/du)\operatorname{dn}(u,k')$ and referring to [23, p. 511], we obtain the expansion

$$D_{11}|_{\theta_1=\pi}=\frac{\psi_1^{3/2}}{2\pi a}\left(\frac{K}{K'}\right)^2\sum_{n=1}^{\infty}\frac{nq'^n}{1+q'^{2n}}\sin\frac{nK\eta_1}{K'}, \tag{7.3}$$

where $q'=\exp(-\pi K/K')$. Denoting the remainder after the first term of the series in (7.3) by R, we have

$$|R|<\sum_{n=2}^{\infty}nq'^n<\frac{2q'^2}{(1-q')^2} \tag{7.4}$$

and hence,

$$D_{11}|_{\theta_1=\pi}=\frac{\psi_1^{3/2}}{2\pi a}\left(\frac{K}{K'}\right)^2q'\sin\frac{K\eta_1}{K'}[1+O(q')] \tag{7.5}$$

as $\varepsilon\to 0$. Finally, expressing (7.5) in terms of geometric quantities by use of (5.7) and the asymptotic results of § 2, we obtain

$$D_{11}|_{\theta_1=\pi}=(\pi/\sigma)\sin\beta\,(\sigma/\varepsilon)^{3/2}\exp[-2^{-1}\pi^2(\sigma/\varepsilon)^{1/2}][1+O(\varepsilon)] \tag{7.6}$$

where

$$\beta=\frac{\pi r_2}{r_1+r_2},\qquad \sigma=\frac{2r_1r_2}{r_1+r_2}. \tag{7.7}$$

When the radii of the spheres are equal, we have $\eta_1=-\frac{1}{2}\log q=\pi K'/2K$, and we find

$$D_{11}|_{\theta_1=\pi}=\frac{(1+q^{1/2})^2}{2\pi^3(1-q^{1/2})q^{1/4}}\cdot\frac{K^2k'^2k^{1/2}}{1+k}. \tag{7.8}$$

Transforming (7.8) to the modulus k_1 by Landen's transformation [23, p. 507], where $k_1=(1-k')/(1+k')$, and denoting the corresponding values of q and K by q_1 and K_1, respectively, we obtain

$$D_{11}|_{\theta_1=\pi}=\frac{(1+q_1)^2}{4\pi^3(1-q_1)q_1^{1/2}}\cdot k_1k_1'K_1^2. \tag{7.9}$$

Equation (7.9) agrees with [10, p. 100] when the subscripts are omitted, after correcting 2π to 4π on pp. 89 and 97 of [10].

8. Behavior of D_1^* at the inner axial point. When $\theta_1=\pi$, we have $\gamma_1=\pi-\theta_0$. Substituting $\theta_1=\pi$ and $\eta_2=\eta_1-\pi K'/K$ in (6.9), replacing t by $\pi-t$, and using Jacobi's imaginary transformation, we obtain

$$\begin{aligned}D_1^*|_{\theta_1=\pi}=&-\frac{K^2k'^2}{4\pi^4a^2}\psi_0^{1/2}\psi_1^{3/2}\int_0^{\theta_0}(2\cos t-2\cos\theta_0)^{-1/2}\\&\cdot\operatorname{Re}\operatorname{sn}\left[\frac{K}{\pi}(\eta_1-\eta_0+it),k'\right]\operatorname{cn}\left[\frac{K}{\pi}(\eta_1-\eta_0+it),k'\right]dt.\end{aligned} \tag{8.1}$$

Expanding the integrand as in (7.3) and interchanging summation and integration, we find

(8.2)

$$D_1^*|_{\theta_1=\pi} = -\frac{\psi_0^{1/2}\psi_1^{3/2}}{2\pi^2 a^2}\left(\frac{K}{K'}\right)^2 \sum_{n=1}^{\infty} \frac{nq'^n}{1+q'^{2n}} \sin\frac{nK}{K'}(\eta_1-\eta_0)\int_0^{\theta_0} \frac{\cosh\dfrac{nK}{K'}\,t\,dt}{(2\cos t - 2\cos\theta_0)^{1/2}}.$$

When $0 \leqq \theta_0 < \pi$, the interchange of summation and integration is justified by [3, p. 495]. Referring to [23, p. 315], we can write (8.2) as a series of Legendre functions of the form

$$D_1^*|_{\theta_1=\pi} = -\frac{\psi_0^{1/2}\psi_1^{3/2}}{4\pi a^2}\left(\frac{K}{K'}\right)^2 \sum_{n=1}^{\infty} \frac{nq'^n}{1+q'^{2n}} \sin\frac{nK}{K'}(\eta_1-\eta_0) P_{-1/2+i(nK/K')}(\cos\theta_0). \tag{8.3}$$

As $\varepsilon \to 0$, we will assume that the point (x_0, ρ_0, ϕ_0) is fixed and that it remains outside both spheres. Defining

$$g = \lim_{\varepsilon\to 0} \frac{\eta_0}{\eta_1} = \frac{2x_0 r_1}{x_0^2+\rho_0^2} \tag{8.4}$$

we see that the condition

$$-\frac{r_2}{r_1} < g < 1 \tag{8.5}$$

insures that the point charge lies outside both spheres when they are tangent to each other.

We have

$$\left| \int_0^{\theta_0} \frac{\cosh\dfrac{nKt}{K'}\,dt}{(2\cos t - 2\cos\theta_0)^{1/2}} \right| < \pi \cosh\frac{nK\theta_0}{K'} < \pi q'^{-n\theta_0/\pi} \tag{8.6}$$

for θ_0 sufficiently near 0. Hence, writing

$$D_1^*|_{\theta_1=\pi} = -\frac{\psi_0^{1/2}\psi_1^{3/2}}{4\pi a^2}\left(\frac{K}{K'}\right)^2 \left\{ \frac{q' \sin\dfrac{K}{K'}(\eta_1-\eta_0)}{1+q'^2} P_{-1/2+i(K/K')}(\cos\theta_0) + R \right\}, \tag{8.7}$$

we obtain

$$|R| < \pi \sum_{n=2}^{\infty} nq'^{n(1-\theta_0/\pi)} < \frac{2\pi q'^{2(1-\theta_0/\pi)}}{(1-q'^{1-\theta_0/\pi})^2}. \tag{8.8}$$

Since $\theta_0 \to 0$ as $\varepsilon \to 0$, we have $|R| = O(q'^{1-\xi})$ for any fixed number ξ in the interval $0 < \xi < 1$.

When $0 \leqq \theta_0 < \sqrt{6}$, we have

$$\theta_0^2 - t^2 > 2\cos t - 2\cos\theta_0 > \left(1-\frac{\theta_0^2}{6}\right)(\theta_0^2-t^2) > 0, \tag{8.9}$$

and hence,

$$(8.10)\quad \int_0^1 \frac{\cosh\frac{K\theta_0 t}{K'}\,t}{\sqrt{1-t^2}}\,dt < \int_0^{\theta_0} \frac{\cosh\frac{K}{K'}\,t\,dt}{(2\cos t - 2\cos\theta_0)^{1/2}} < \frac{1}{\sqrt{1-\frac{\theta_0^2}{6}}}\int_0^1 \frac{\cosh\frac{K\theta_0}{K'}\,t}{\sqrt{1-t^2}}\,dt.$$

Referring to [23, p. 366], we find that (8.10) is equivalent to

$$(8.11)\quad I_0\left(\frac{K\theta_0}{K'}\right) < P_{-1/2+i(K/K')}(\cos\theta_0) < \frac{1}{\sqrt{1-\frac{\theta_0^2}{6}}}\,I_0\left(\frac{K\theta_0}{K'}\right),$$

where $I_0(z)$ is the modified Bessel function of the first kind of order zero. It follows that

$$(8.12)\quad P_{-1/2+i(K/K')}(\cos\theta_0) = I_0\left(\frac{K\theta_0}{K'}\right)[1+O(\varepsilon)],$$

as $\varepsilon \to 0$. Finally, from (8.7), (8.12), and the asymptotic results of § 2, we have[1]

$$(8.13)\quad D_1^*|_{\theta_1=\pi} = -\frac{(\pi/\sigma)}{\sqrt{x_0^2+\rho_0^2}}\sin\lambda\, I_0(\beta h)(\sigma/\varepsilon)^{3/2}\exp\left[-2^{-1}\pi^2(\sigma/\varepsilon)^{1/2}\right][1+O(\varepsilon)]$$

where β and σ are given by (7.8), g is defined by (8.4), and

$$(8.14)\quad h = \lim_{\varepsilon\to 0}\frac{\theta_0}{\eta_1} = \frac{2\rho_0 r_1}{x_0^2+\rho_0^2}, \qquad \lambda = \beta(1-g).$$

The preceding results do not hold uniformly with respect to the location of the point charge when $0 \leqq \theta_0 \leqq \pi$. When $x_0 = 0$, we see that the value of ε required for a given relative error tends to zero as ρ_0 tends to zero.

9. An asymptotic expansion for D_{12} at the inner axial point. Substituting $\eta_2 = \eta_1 - \pi K'/K$ in (6.4) and using the addition theorems for sn u and cn u, we find that

$$(9.1)\quad D_{12} = \frac{K^2}{2\pi^4 a}\psi_1^{3/2}\int_0^{\theta_1}(2\cos t - 2\cos\theta_1)^{-1/2}\,\mathrm{Re}\,\frac{\mathrm{dn}\left[\frac{K}{\pi}(\eta_1 - it), k'\right]}{\mathrm{sn}^2\left[\frac{K}{\pi}(\eta_1 - it), k'\right]}\,dt.$$

Hence,

$$(9.2)\quad D_{12}|_{\theta_1=\pi} = \frac{K^2}{4\pi^4 a}\psi_1^{3/2}\int_0^{\pi}\mathrm{Re}\,\frac{\mathrm{dn}\left[\frac{K}{\pi}(\eta_1 - it), k'\right]}{\mathrm{sn}^2\left[\frac{K}{\pi}(\eta_1 - it), k'\right]}\sec\frac{t}{2}\,dt.$$

Noting that dn $u/\mathrm{sn}^2 u = -(d/du)$ cs u, and referring to [23, p. 512], we obtain

$$(9.3)\quad \frac{\mathrm{dn}\left(\frac{Kz}{\pi}, k'\right)}{\mathrm{sn}^2\left(\frac{Kz}{\pi}, k'\right)} = \left(\frac{\pi}{2K'}\right)^2\csc^2\frac{Kz}{2K'} + \frac{2\pi^2}{K'^2}\sum_{n=1}^{\infty}\frac{nq'^{2n}}{1+q'^{2n}}\cos\frac{nKz}{K'}$$

[1] This corrects the corresponding result in the author's PhD. thesis.

for $|\operatorname{Im} z|<2\pi$. In particular, (9.3) holds for $\operatorname{Im} z=-\pi$. Hence, the left-hand side of (9.3) is purely imaginary for $z=\eta_1-i\pi$, and we have

$$\operatorname{Re}\frac{\operatorname{dn}\left[\frac{K}{\pi}(\eta_1-it),k'\right]}{\operatorname{sn}^2\left[\frac{K}{\pi}(\eta_1-it),k'\right]}=\frac{1}{2}\left(\frac{\pi}{K'}\right)^2\left\{\frac{1-\cos\frac{K\eta_1}{K'}\cosh\frac{Kt}{K'}}{\left(\cosh\frac{Kt}{K'}-\cos\frac{K\eta_1}{K'}\right)^2}-\frac{1-\cos\frac{K\eta_1}{K'}\cosh\frac{K\pi}{K'}}{\left(\cosh\frac{K\pi}{K'}-\cos\frac{K\eta_1}{K'}\right)^2}\right\}$$

$$+2\left(\frac{\pi}{K'}\right)^2\sum_{n=1}^{\infty}\frac{nq'^{2n}\cos\frac{nK\eta_1}{K'}}{1+q'^{2n}}\left(\cosh\frac{nKt}{K'}-\cosh\frac{nK\pi}{K'}\right). \tag{9.4}$$

From (9.2) and (9.4), justifying interchange of summation and integration by the Weierstrass test, we obtain

$$D_{12}|_{\theta_1=\pi}=\frac{\psi_1^{3/2}}{8\pi^2 a}\left(\frac{K}{K'}\right)^2\left\{\int_0^{\pi}\left[\frac{1-\cos\frac{K\eta_1}{K'}\cosh\frac{Kt}{K'}}{\left(\cosh\frac{Kt}{K'}-\cos\frac{K\eta_1}{K'}\right)^2}-\frac{1-\cos\frac{K\eta_1}{K'}\cosh\frac{K\pi}{K'}}{\left(\cosh\frac{K\pi}{K'}-\cos\frac{K\eta_1}{K'}\right)^2}\right]\sec\frac{t}{2}\,dt\right.$$

$$\left.+4\sum_{n=1}^{\infty}\frac{nq'^{2n}}{1+q'^{n}}\cos\frac{nK\eta_1}{K'}\int_0^{\pi}\left(\cosh\frac{nKt}{K'}-\cosh\frac{nK\pi}{K'}\right)\sec\frac{t}{2}\,dt\right\}. \tag{9.5}$$

Denoting the sum in (9.5) by S and the integral in the summation by I, we have

$$\begin{aligned}|I|&=2\int_0^{\pi}\frac{\sinh\frac{nK(\pi+t)}{2K'}\sinh\frac{nK(\pi-t)}{2K'}}{\cos(t/2)}\,dt\\&<2\pi\sinh\frac{nK\pi}{K'}\int_0^{nK\pi/2K'}\frac{\sinh t}{t}\,dt\\&<2\pi\sinh\frac{nK\pi}{K'}\cosh\frac{nK\pi}{2K'}\\&<\pi q'^{-3n/2}\end{aligned} \tag{9.6}$$

and

$$|S|<\pi\sum_{n=1}^{\infty}nq'^{n/2}=\pi\frac{q'^{1/2}}{(1-q'^{1/2})^2}. \tag{9.7}$$

Hence, with $\delta=\eta_1-\eta_2=\pi K'/K$ as before and $\mu=K\eta_1/K'$, we have

$$D_{12}|_{\theta_1=\pi}=\frac{\psi_1^{3/2}}{8\pi a\delta}\left\{\int_0^{\pi^2/\delta}[g(u)-g(\pi^2/\delta)]\sec\frac{\delta u}{2\pi}\,du+R^{(1)}\right\}, \tag{9.8}$$

where

$$g(u)=\frac{1-\cos\mu\cosh u}{(\cosh u-\cos\mu)^2}, \tag{9.9}$$

and

$$|R^{(1)}|<\frac{4\pi q'^{1/2}}{(1-q'^{1/2})^2}. \tag{9.10}$$

We have the identity

$$\sec x = \sum_{m=0}^{n} \frac{(-1)^m E_{2m}}{(2m)!} x^{2m} + \frac{(-4)^{n+1} x^{2n+2}}{(2n+1)! \cos x} \int_0^{1/2} E_{2n+1}(t) \cos 2xt \, dt \tag{9.11}$$

[15, p. 35], where the E_m and $E_m(t)$ are the Euler numbers and Euler polynomials, respectively. Noting that $(-1)^{n+1} E_{2n+1}(t) \geqq 0$ for $0 \leqq t \leqq \frac{1}{2}$, and referring to [15, pp. 24–25], we obtain

$$(-1)^{n+1} \int_0^{1/2} E_{2n+1}(t) \cos 2xt \, dt < (-1)^{n+1} \int_0^{1/2} E_{2n+1}(t) \, dt = \frac{(-1)^{n+1} E_{2n+2}}{(2n+2) 2^{2n+2}} \tag{9.12}$$

when $0 < x < \pi/2$. We therefore have

$$\begin{aligned} &\int_0^{\pi^2/\delta} [g(u) - g(\pi^2/\delta)] \sec \frac{\delta u}{2\pi} \, du \\ &\quad = \sum_{m=0}^{n} \frac{(-1)^m E_{2m}}{(2m)!} \left(\frac{\delta}{2\pi}\right)^{2m} \int_0^{\pi^2/\delta} u^{2m} [g(u) - g(\pi^2/\delta)] \, du + R_n^{(2)}, \end{aligned} \tag{9.13}$$

where

$$|R_n^{(2)}| < \frac{(-1)^{n+1} E_{2n+2}}{(2n+2)!} \left(\frac{\delta}{2\pi}\right)^{2n+2} \int_0^{\pi^2/\delta} u^{2n+2} |g(u) - g(\pi^2/\delta)| \sec \frac{\delta u}{2\pi} \, du. \tag{9.14}$$

We have $\mu = \pi\eta_1/(\eta_1 - \eta_2)$, and hence, μ takes the values $0 \leqq \mu < \pi$ when $0 \leqq \eta_1 < \infty$. When $0 \leqq \mu \leqq \pi$, we find that

$$\begin{aligned} |g(u)| &< \frac{1 + \cosh u}{(\cosh u - 1)^2} = \frac{2(1 + e^{-u})^2}{(1 - e^{-u})^4} e^{-u} \\ &< 7.04 e^{-u} \end{aligned} \tag{9.15}$$

when $u > \log 5$, and

$$|g(u)| < 2.0001 e^{-u} \tag{9.16}$$

when $u > 12$. Similarly, from

$$g'(u) = -\frac{2 \sin^2 \mu \sinh u}{(\cosh u - \cos \mu)^3} + \frac{\cos \mu \sinh u}{(\cosh u - \cos \mu)^2}, \tag{9.17}$$

we obtain

$$\begin{aligned} |g'(u)| &< \frac{2 \sinh u}{(\cosh u - 1)^3} + \frac{\sinh u}{(\cosh u - 1)^2} \\ &= \left[\frac{8(1 + e^{-u}) e^{-u}}{(1 - e^{-u})^5} + \frac{2(1 + e^{-u})}{(1 - e^{-u})^3}\right] e^{-u} \\ &< 2.0001 e^{-u} \end{aligned} \tag{9.18}$$

when $0 \leqq \mu \leqq \pi$ and $u > 12$. We also have

$$u^2 |g(u)| < \frac{u^2 (1 - \cos \mu)}{(1 - \cos \mu + u^2/2)^2} < 1 \tag{9.19}$$

for all $u > 0$ and for $0 \leqq \mu \leqq \pi$, and

$$x_n e^{-x} \leqq \int_x^{\infty} u^n e^{-u} \, du < 2x^n e^{-x} \tag{9.20}$$

for $x>2n$, $n\geqq 0$. The second inequality follows from the asymptotic expansion for the incomplete gamma function [1, p. 263], and can also be obtained by use of integration by parts and induction. From the preceding, we find that

$$\begin{aligned}
&\int_0^{\pi^2/2\delta} u^{2n+2}|g(u)-g(\pi^2/\delta)|\sec\frac{\delta u}{2\pi}\,du \\
&\quad<\sqrt{2}\int_0^{\log 5} u^{2n}\,du+7.04\sqrt{2}\int_{\log 5}^{\pi^2/2\delta} u^{2n+2}\,e^{-u}\,du+\frac{7.04\sqrt{2}}{2n+3}\left(\frac{\pi^2}{2\delta}\right)^{2n+3} e^{-\pi^2/\delta} \\
&\quad=\sqrt{2}\int_0^{\log 5}[u^{2n}-7.04u^{2n+2}\,e^{-u}]\,du \qquad (9.21)\\
&\qquad+7.04\sqrt{2}\int_0^{\pi^2/2\delta} u^{2n+2}\,e^{-u}\,du+\frac{7.04\sqrt{2}}{2n+3}\left(\frac{\pi^2}{2\delta}\right)^{2n+3} e^{-\pi^2/\delta} \\
&\quad<7.04\sqrt{2}\left\{(2n+2)!-(\pi^2/2\delta)^{2n+2}\,e^{-\pi^2/2\delta}\left[1-\frac{e^{-1}}{2n+3}\right]\right\} \\
&\quad<7.04\sqrt{2}(2n+2)!
\end{aligned}$$

for $0\leqq\mu\leqq\pi$ and $\pi^2/\delta>4n+4$, $n\geqq 0$, since $\log 5<2$, $(\pi^2/2\delta)^{2n+2}\,e^{-\pi^2/2\delta}<(2n+2)!/2\sqrt{\pi}$, and $\int_0^{\log 5}(u^{2n}-7.04u^{2n+2}\,e^{-u})\,du<\int_0^{\log 5}(u^{2n}-7.04u^{2n+2}/5)\,du<0$. Noting that $|g(u)-g(\pi^2/\delta)|'\leqq|g'(u)|$, we find from (9.18) that

$$\begin{aligned}
\left|g(u)-g\left(\frac{\pi^2}{\delta}\right)\right|\sec\frac{\delta u}{2\pi}&<2.0001(e^{-u}-e^{-\pi^2/\delta})\sec\frac{\delta u}{2\pi} \\
&<\frac{4.0002\pi}{\delta}\left(\frac{e^{-u}-e^{-\pi^2/\delta}}{\pi^2/\delta-u}\right) \qquad (9.22)\\
&<\frac{4.0002\pi}{\delta}\,e^{-u}
\end{aligned}$$

for $u>12$. Hence,

$$\begin{aligned}
\int_{\pi^2/2\delta}^{\pi^2/\delta} u^{2n+2}|g(u)-g(\pi^2/\delta)|\sec\frac{\delta u}{2\pi}\,du&<\frac{8.0004}{\pi}\left(\frac{\pi^2}{2\delta}\right)\int_{\pi^2/2\delta}^{\pi^2/\delta} u^{2n+2}\,e^{-u}\,du \qquad (9.23)\\
&<\frac{16.0008}{\pi}\left(\frac{\pi^2}{2\delta}\right)^{2n+3} e^{-\pi^2/2\delta}
\end{aligned}$$

when $\pi^2/\delta>24$. It follows from (9.21) and (9.23) that $R_n^{(2)}=O(\delta^{2n+2})$ as $\delta\to 0$, and hence, that the right-hand side of (9.13) is an asymptotic expansion. We see that (9.13) is also convergent, since it can be shown that integration and summation can be interchanged when the Maclaurin series for $\sec(\delta u/2\pi)$ is substituted in the left side of (9.13). Alternatively, it can be shown directly that $R_n^{(2)}$ is $O(n^{-1})$ as $n\to\infty$ and that the nth term of the series in (9.13) is $O(n^{-2})$.

To obtain a simpler, but divergent, asymptotic expansion, we write

$$\begin{aligned}
&\sum_{m=0}^{n}\frac{(-1)^m E_{2m}}{(2m)!}\left(\frac{\delta}{2\pi}\right)^{2m}\int_0^{\pi^2/\delta} u^{2m}[g(u)-g(\pi^2/\delta)]\,du \qquad (9.24)\\
&\quad=\sum_{m=0}^{n}\frac{(-1)^m E_{2m}}{(2m)!}\left(\frac{\delta}{2\pi}\right)^{2m}\int_0^{\infty} u^{2m}g(u)\,du+R_n^{(3)},
\end{aligned}$$

where

$$R_n^{(3)} = -\sum_{m=0}^{n} \frac{(-1)^m E_{2m}}{(2m)!}\left(\frac{\delta}{2\pi}\right)^{2m}\left\{\int_{\pi^2/\delta}^{\infty} u^{2m} g(u)\, du + \frac{g(\pi^2/\delta)}{2m+1}\left(\frac{\pi^2}{\delta}\right)^{2m+1}\right\}. \tag{9.25}$$

It follows from (9.16) and (9.20) that

$$|R_n^{(3)}| < \sum_{m=0}^{n} \frac{(-1)^m E_{2m}}{(2m)!}\left(\frac{\pi}{2}\right)^{2m}\left\{2\, e^{-\pi^2/\delta} + \frac{2.0001}{2m+1}\left(\frac{\pi^2}{\delta}\right) e^{-\pi^2/\delta}\right\} \tag{9.26}$$

for $\pi^2/\delta > 12$.

Referring to [4, p. 38], we have

$$\begin{aligned}\int_0^{\infty} u^{2m} g(u)\, du &= -\frac{d}{d\mu}\sin\mu \int_0^{\infty} \frac{u^{2m}\, du}{\cosh u - \cos\mu} \\ &= (-1)^m (2\pi)^{2m} B_{2m}\left(\frac{\mu}{2\pi}\right),\end{aligned} \tag{9.27}$$

where $B_n(x)$ is the nth Bernoulli polynomial. Finally, noting that $\delta = -\log q$, we obtain the asymptotic expansion

$$D_{12}|_{\theta_1=\pi} = -\frac{\psi_1^{3/2}}{8\pi a \log q}\left\{\sum_{m=0}^{n} \frac{E_{2m} B_{2m}(\mu/2\pi)}{(2m)!}(\log q)^{2m} + R_n\right\} \tag{9.28}$$

for $|\log q| < \pi^2/(4n+12)$, $n \geqq 0$, where

$$R_n = R^{(1)} + R_n^{(2)} + R_n^{(3)}. \tag{9.29}$$

When δ is sufficiently small, we can obtain more convenient bounds for the exponentially small terms in the remainder. By use of the inequality

$$(2n+2)! > (2n+2)^{2n+2}\, e^{-(2n+2)}\sqrt{2\pi(2n+2)}, \tag{9.30}$$

$n \geqq 0$, we obtain

$$\frac{x^{2n+3}\, e^{-x}}{(2n+2)!} < \left(\frac{x}{2n+2}\right)^{2n+3} e^{-x+2n+2}\sqrt{\frac{n+1}{\pi}} \tag{9.31}$$

for $x > 0$. We see that $x^{2n+3}\, e^{-x}$ is decreasing for $x > 2n+3$. Hence, when $x > 2c(n+1)$, $c > 1$, we have

$$\begin{aligned}\frac{x^{2n+3}\, e^{-x}}{(2n+2)!} &< c^{2n+3}\, e^{-(c-1)(2n+2)}\sqrt{\frac{n+1}{\pi}} \\ &< \frac{c^3\, e^{-2(c-1)}}{\sqrt{\pi}},\end{aligned} \tag{9.32}$$

since the right side of the first line of (9.32) is decreasing with respect to n for $n \geqq 0$. In particular, we find

$$\frac{x^{2n+3}\, e^{-x}}{(2n+2)!} < \frac{3^3\, e^{-4}}{\sqrt{\pi}} < 0.28, \qquad x > 6n+6 \tag{9.33}$$

$$< \frac{6^3\, e^{-10}}{\sqrt{\pi}} < 0.0056, \qquad x > 12n+12. \tag{9.34}$$

Similarly, we have

$$\frac{x^{2n+2}\, e^{-x}}{(2n+2)!} < \frac{6^2\, e^{-10}}{2\sqrt{\pi}} < 0.0005 \tag{9.35}$$

and

$$\frac{x^{2n+2}\,e^{-x/2}}{(2n+2)!}<\frac{6^2\,e^{-4}}{2\sqrt{\pi}}<0.1861 \tag{9.36}$$

when $x>12n+12$.

From (9.10) and (9.36), we obtain

$$|R^{(1)}|<2.36(2n+2)!\left(\frac{\delta}{\pi^2}\right)^{2n+2} \tag{9.37}$$

when $\pi^2/\delta>12n+12$. Similarly, from (9.14), (9.21), (9.23), and (9.33), it follows that

$$|R_n^{(2)}|<11.4(-1)^{n+1}E_{2n+2}\left(\frac{\pi}{2}\right)^{2n+2}\left(\frac{\delta}{\pi^2}\right)^{2n+2}. \tag{9.38}$$

Using the inequality

$$\frac{(-1)^nE_{2n}}{(2n)!}\left(\frac{\pi}{2}\right)^{2n}<\frac{4}{\pi} \tag{9.39}$$

[1, p. 805], we obtain the simpler bounds

$$|R_n^{(2)}|<14.6(2n+2)!\left(\frac{\delta}{\pi^2}\right)^{2n+2}, \tag{9.40}$$

$\pi^2/\delta>12n+12$, and

$$|R_n^{(3)}|<\frac{8}{\pi}(n+1)\,e^{-\pi^2/\delta}+\frac{8.0004}{\pi}\left(\frac{\pi^2}{\delta}\right)e^{-\pi^2/\delta}\sum_{m=0}^{n}1/(2m+1), \tag{9.41}$$

$\pi^2/\delta>12$. As in (9.31)-(9.36), we find that $(n+1)x^{2n+2}\,e^{-x}$ and $x^{2n+3}\,e^{-x}\sum_{m=0}^{n}1/(2m+1)$ have the same upper bounds as (9.35) and (9.34), respectively, when $x>12n+12$. We therefore obtain

$$|R_n^{(3)}|<0.016(2n+2)!\left(\frac{\delta}{\pi^2}\right)^{2n+2} \tag{9.42}$$

when $\pi^2/\delta>12n+12$, which is almost negligible compared with $R^{(1)}$ and $R_n^{(2)}$. Finally, we have

$$|R_n|<17.0(2n+2)!\left(\frac{\delta}{\pi^2}\right)^{2n+2} \tag{9.43}$$

for $\pi^2/\delta>12n+12$ and $0\leqq\mu\leqq\pi$.

We see that μ is a function of ε for given values of r_1 and r_2 except in the cases $r_1\to\infty$ with r_2 constant, $r_1=r_2$, and $r_2\to\infty$ with r_1 constant. We then have $\mu=0$, $\pi/2$, and π, respectively. In the case $\mu=0$, sphere 1 is an infinite plane when r_2 is finite. Similarly, when $\mu=\pi$, sphere 2 is an infinite plane when r_1 is finite. We have

$$\begin{aligned} B_{2n}(\mu/2\pi)&=B_{2n}, && \mu=0\\ &=-2^{-2n}(1-2^{1-2n})B_{2n}, && \mu=\pi/2\\ &=-(1-2^{1-2n})B_{2n}, && \mu=\pi \end{aligned} \tag{9.44}$$

[1, pp. 805-806], and from (2.9)-(2.12),

$$
\begin{aligned}
-\log q &= \sinh^{-1}\frac{a}{r_2} \sim \sqrt{\frac{2\varepsilon}{r_2}}, && \mu = 0 \\
&= 2\sinh^{-1}\frac{a}{r} \sim 2\sqrt{\frac{\varepsilon}{r}}, && r_1 = r_2 = r \\
&= \sinh^{-1}\frac{a}{r_1} \sim \sqrt{\frac{2\varepsilon}{r_1}}, && \mu = \pi.
\end{aligned} \tag{9.45}
$$

We can express (9.28) as an asymptotic expansion in powers of ε by use of (2.7)-(2.9). The approximation of lowest order is given by

$$D_{12}|_{\theta_1=\pi} = (2\pi\varepsilon)^{-1}[1+O(\varepsilon)], \tag{9.46}$$

and is independent of the radii. The leading term of (9.46) is also given by an approximate analysis of Maxwell [12, p. 154] for the case of two nearly parallel charged surfaces.

10. Integral representations for the limiting charge density as $\varepsilon \to 0$. It follows from (7.1), (9.1), and (9.3) that

$$
\begin{aligned}
D_{11} = \frac{\psi_1^{3/2}}{4\pi^2 a}\left(\frac{K}{K'}\right)^2 \Bigg\{ & \sin\frac{K\eta_1}{K'} \int_{\theta_1}^{\pi} \frac{\sinh\frac{Kt}{K'}\,dt}{\left(\cosh\frac{Kt}{K'} - \cos\frac{K\eta_1}{K'}\right)^2 (2\cos\theta_1 - 2\cos t)^{1/2}} \\
& + 4\sum_{n=1}^{\infty} \frac{nq'^{2n}\sin\frac{nK\eta_1}{K'}}{1+q'^{2n}} \int_{\theta_1}^{\pi} \frac{\sinh\frac{nKt}{K'}\,dt}{(2\cos\theta_1 - 2\cos t)^{1/2}} \Bigg\}
\end{aligned} \tag{10.1}
$$

and

$$
\begin{aligned}
D_{12} = \frac{\psi_1^{3/2}}{4\pi^2 a}\left(\frac{K}{K'}\right)^2 \Bigg\{ & \int_0^{\theta_1} \frac{\left(1 - \cos\frac{K\eta_1}{K'}\cosh\frac{Kt}{K'}\right) dt}{\left(\cosh\frac{Kt}{K'} - \cos\frac{K\eta_1}{K'}\right)^2 (2\cos t - 2\cos\theta_1)^{1/2}} \\
& + 4\sum_{n=1}^{\infty} \frac{nq'^{2n}\cos\frac{nK\eta_1}{K'}}{1+q'^{2n}} \int_0^{\theta_1} \frac{\cosh\frac{nKt}{K'}\,dt}{(2\cos t - 2\cos\theta_1)^{1/2}} \Bigg\}
\end{aligned} \tag{10.2}
$$

when $0 < \theta_1 < \pi$. Let the sums in (10.1) and (10.2) be denoted by S_1 and S_2, respectively. We have

$$
\begin{aligned}
\left| \int_{\theta_1}^{\pi} \frac{\sinh\frac{nKt}{K'}\,dt}{(2\cos\theta_1 - 2\cos t)^{1/2}} \right| &< \sinh\frac{nK\pi}{K'} \int_{\theta_1}^{\pi} \frac{dt}{(2\cos\theta_1 - 2\cos t)^{1/2}} \\
&< \frac{1}{2} q'^{-n} K(\cos\theta_1)
\end{aligned} \tag{10.3}
$$

where $K(k)$ is the complete elliptic integral of the first kind with modulus k. Hence,

$$
\begin{aligned}
|S_1| &< \frac{1}{2} K(\cos\theta_1) \sum_{n=1}^{\infty} nq'^n = \frac{q'K(\cos\theta_1)}{2(1-q')^2} \\
&= O[q'\log(1-q)]
\end{aligned} \tag{10.4}
$$

as $\varepsilon \to 0$. Similarly,

$$\left| \int_0^{\theta_1} \frac{\cosh \frac{nKt}{K'}\, dt}{(2\cos t - 2\cos\theta_1)^{1/2}} \right| < q'^{-n} K(\sin\theta_1), \tag{10.5}$$

and

$$\begin{aligned} |S_2| &< \frac{q'K(\sin\theta_1)}{(1-q')^2} \\ &= O(q'). \end{aligned} \tag{10.6}$$

It follows that

$$\lim_{\varepsilon\to 0} D_{11} = \lim_{\varepsilon\to 0} \frac{\psi_1^{3/2}}{4\pi^2 a} \left(\frac{K}{K'}\right)^2 \sin\frac{K\eta_1}{K'}\, I_1 \tag{10.7}$$

and

$$\lim_{\varepsilon\to 0} D_{12} = \lim_{\varepsilon\to 0} \frac{\psi_1^{3/2}}{4\pi^2 a} \left(\frac{K}{K'}\right)^2 I_2, \tag{10.8}$$

where

$$I_1 = \int_{\theta_1}^{\pi} \frac{\sinh \frac{Kt}{K'}\, dt}{\left(\cosh\frac{Kt}{K'} - \cos\frac{K\eta_1}{K'}\right)^2 (2\cos\theta_1 - 2\cos t)^{1/2}}, \tag{10.9}$$

$$I_2 = \int_0^{\theta_1} \frac{\left(1 - \cos\frac{K\eta_1}{K'}\cosh\frac{Kt}{K'}\right) dt}{\left(\cosh\frac{Kt}{K'} - \cos\frac{K\eta_1}{K'}\right)^2 (2\cos t - 2\cos\theta_1)^{1/2}}. \tag{10.10}$$

Substituting $\sin(t/2) = u\sin(\theta_1/2)$ in (10.9) and (10.10), we obtain

$$I_1 = \int_1^{\csc(\theta_1/2)} \frac{\sinh \frac{Kt(u)}{K'}\, du}{\left[\cosh\frac{Kt(u)}{K'} - \cos\frac{K\eta_1}{K'}\right]^2 \sqrt{u^2-1}} \tag{10.11}$$

and

$$I_2 = \int_0^1 \frac{\left(1 - \cos\frac{K\eta_1}{K'}\cosh\frac{Kt(u)}{K'}\right) du}{\left[\cosh\frac{Kt(u)}{K'} - \cos\frac{K\eta_1}{K'}\right]^2 \sqrt{1-u^2}}, \tag{10.12}$$

where $t(u) = 2\sin^{-1}(u\sin(\theta_1/2))$. We have

$$\lim_{\varepsilon\to 0} \frac{K\eta_1}{K'} = \beta, \tag{10.13}$$

where β is given by (7.7), and

$$\lim_{\varepsilon\to 0} \frac{Kt(u)}{K'} = \lim_{\varepsilon\to 0} \frac{K\theta_1 u}{K'} = \alpha u, \tag{10.14}$$

where

$$\alpha = \beta \cot(\omega_1/2). \tag{10.15}$$

We also have

$$\lim_{\varepsilon\to 0} \frac{\psi_1^{3/2}}{4\pi^2 a}\left(\frac{K}{K'}\right)^2 = \frac{r_2^2 \csc^3(\omega_1/2)}{4r_1(r_1+r_2)^2}. \tag{10.16}$$

It can be shown by the Weierstrass test that I_1 and I_2 are uniformly convergent with respect to ε in an interval $0 \leqq \varepsilon \leqq \varepsilon_1$ for $0 < \omega_1 \leqq \pi$. Hence, they are continuous at $\varepsilon = 0$. Finally, when $\omega_1 \neq 0$, we obtain

$$\lim_{\varepsilon\to 0} D_{11} = \frac{r_2^2 \csc^3(\omega_1/2)}{4r_1(r_1+r_2)^2} \sin\beta \int_1^\infty \frac{\sinh \alpha u\, du}{(\cosh \alpha u - \cos\beta)^2\sqrt{u^2-1}} \tag{10.17}$$

and

$$\lim_{\varepsilon\to 0} D_{12} = \frac{r_2^2 \csc^3(\omega_1/2)}{4r_1(r_1+r_2)^2} \int_0^1 \frac{(1-\cos\beta\cosh\alpha u)\, du}{(\cosh\alpha u - \cos\beta)^2\sqrt{1-u^2}}. \tag{10.18}$$

The discussion for D_{11} holds also for D_1^* when we assume that (8.5) is satisfied, giving

$$\lim_{\varepsilon\to 0} D_1^* = -\lim_{\varepsilon\to 0} \frac{\psi_0^{1/2}\psi_1^{3/2}}{8\pi^2 a^2}\left(\frac{K}{K'}\right)^2 \sin\frac{K(\eta_1-\eta_0)}{K'} I_3, \tag{10.19}$$

where

$$I_3 = \int_{\gamma_1}^{\pi} \frac{\sinh\dfrac{Kt}{K'}\, dt}{\left(\cosh\dfrac{Kt}{K'} - \cos\dfrac{K(\eta_1-\eta_0)}{K'}\right)^2 (2\cos\gamma_1 - 2\cos t)^{1/2}}. \tag{10.20}$$

Proceeding as before, we find

$$\lim_{\varepsilon\to 0} D_1^* = -\frac{r_2^2 \csc^3(\omega_1/2)}{4r_1(r_1+r_2)^2(x_0^2+\rho_0^2)^{1/2}} \sin\lambda \int_1^\infty \frac{\sinh\kappa u\, du}{(\cosh\kappa u - \cos\lambda)^2\sqrt{u^2-1}}, \tag{10.21}$$

where

$$\begin{aligned} \kappa &= \lim_{\varepsilon\to 0} K\gamma_1/K' \\ &= \beta\sqrt{\cot^2(\omega_1/2) - 2h\cot(\omega_1/2)\cos(\phi-\phi_0)+h^2}, \end{aligned} \tag{10.22}$$

and where h and λ are given by (8.14).

We see that (10.17) gives the charge density in the first problem when the spheres are tangent and at unit potential, and that (10.21) gives the charge density in the second problem when the spheres are tangent. Carl Neumann [13, p. 46] has expressed the latter in terms of a definite integral involving the Bessel function $J_0(z)$.

11. Expansions near $\omega_1 = 0$ for the limits of D_{11} and D_1^*. We have

$$\begin{aligned} \frac{\sinh\alpha u}{(\cosh\alpha u - \cos\beta)^2} &= -\frac{1}{u\sin\beta}\frac{\partial^2}{\partial\alpha\,\partial\beta}\log(1-e^{-\alpha u+i\beta})(1-e^{-\alpha u-i\beta}) \\ &= \frac{2}{\sin\beta}\sum_{n=1}^{\infty} n\sin n\beta\, e^{-n\alpha u} \end{aligned} \tag{11.1}$$

for $\alpha u > 0$ when β is real. Substituting (11.1) in (10.17) and referring to [22, p. 172], we find

$$\lim_{\varepsilon\to 0} D_{11} = \frac{r_2^2 \csc^3(\omega_1/2)}{2r_1(r_1+r_2)^2} \sum_{n=1}^{\infty} n \sin n\beta\, K_0(n\alpha), \tag{11.2}$$

where $K_0(z)$ is the modified Bessel function of the second kind of order zero. The interchange of summation and integration is valid by [3, p. 495] as before. From the asymptotic expansion of $K_0(z)$ [23, p. 374], it follows that (11.2) is convergent for $0 < \omega_1 < \pi$. For sufficiently large α, we have

$$\frac{K_0(n\alpha)}{K_0(\alpha)} = \frac{e^{-(n-1)\alpha}}{\sqrt{n}}[1+O(\alpha^{-1})] < e^{-(n-1)\alpha}, \qquad n \geqq 2. \tag{11.3}$$

Hence,

$$\left| \sum_{n=2}^{\infty} n \sin n\beta \frac{K_0(n\alpha)}{K_0(\alpha)} \right| < \sum_{n=1}^{\infty} (n+1)\, e^{-n\alpha} < \frac{2\, e^{-\alpha}}{(1-e^{-\alpha})^2} = O(e^{-\alpha}), \tag{11.4}$$

and

$$\frac{1}{\sin\beta} \sum_{n=1}^{\infty} n \sin n\beta\, K_0(n\alpha) = K_0(\alpha)[1+O(e^{-\alpha})]. \tag{11.5}$$

It follows that as $\omega_1 \to 0$,

$$\begin{aligned} \lim_{\varepsilon\to 0} D_{11} &= \frac{r_2^2 \csc^3(\omega_1/2)}{2r_1(r_1+r_2)^2} \sin\beta\, K_0(\alpha)[1+O(e^{-\alpha})] \\ &= \frac{r_2^2 \csc^3(\omega_1/2) \sin\beta}{2r_1(r_1+r_2)^2} \sqrt{\frac{\pi}{2\alpha}}\, e^{-\alpha} \\ &\quad \cdot \left[1 + \sum_{m=1}^{n} (-1)^m \frac{1^2 \cdot 3^2 \cdots (2m-1)^2}{2^{3m} m!\, \alpha^m} + O(\alpha^{-n-1}) \right]. \end{aligned} \tag{11.6}$$

Expressed in terms of ω_1, we have

$$\lim_{\varepsilon\to 0} D_{11} = \frac{2r_2^2 \sin\beta}{r_1(r_1+r_2)^2} \sqrt{\frac{\pi}{\beta}}\, \omega_1^{-5/2}\, e^{-2\beta/\omega_1}[1+O(\omega_1)]. \tag{11.7}$$

Starting from (10.21), in the same way we find that

$$\lim_{\varepsilon\to 0} D_1^* = -\frac{r_2^2 \csc^3(\omega_1/2)}{2r_1(r_1+r_2)^2(x_0^2+\rho_0^2)^{1/2}} \sum_{n=1}^{\infty} n \sin n\lambda\, K_0(n\kappa) \tag{11.8}$$

for $0 < \lambda < \pi$ and $\kappa > 0$ when (8.5) is satisfied, and that

$$\lim_{\varepsilon\to 0} D_1^* = -\frac{2r_2^2 \sin\lambda}{r_1(r_1+r_2)^2(x_0^2+\rho_0^2)^{1/2}} \sqrt{\frac{\pi}{\beta}}\, \omega_1^{-5/2}\, e^{-2\beta/\omega_1}[1+O(\omega_1)] \tag{11.9}$$

as $\omega_1 \to 0$ with $\omega_1 < 2\cot^{-1} h$. When $\omega_1 = 2\cot^{-1} h$ while $\phi = \phi_0$, we have $\kappa = 0$. We see that the latter occurs arbitrarily close to $\omega_1 = 0$ when $x_0 = 0$ and $\rho_0 > 0$ is sufficiently small.

12. An asymptotic expansion for the limit of D_{12}. When the integrand of

$$I = \int_0^1 \frac{(1-\cos\beta\cosh\alpha u)\, du}{(\cosh\alpha u - \cos\beta)^2 \sqrt{1-u^2}} \tag{12.1}$$

is expanded in powers of $e^{-\alpha u}$ as in § 11 and integration and summation are interchanged, the resulting series does not converge. However, we can obtain an asymptotic

expansion for large α by use of the identity

$$(1-z^2)^{-1/2}=\sum_{m=0}^{n}\frac{(2m)!}{2^{2m}(m!)^2}z^{2m}+f_n(z), \tag{12.2}$$

where

$$f_n(z)=\frac{(2n+1)!}{2^{2n}(n!)^2}(1-z^2)^{-1/2}\int_0^z t^{2n+1}(1-t^2)^{-1/2}\,dt \tag{12.3}$$

[26, p. 108], $n\geqq 0$, which is valid when z^2-1 is not a positive real number. When $0<x<1$, we can verify that

$$\int_0^x t^{2n+1}(1-t^2)^{-1/2}\,dt<\frac{x^{2n+2}}{1+(1-x^2)^{1/2}}<x^{2n+2}, \tag{12.4}$$

and hence, that

$$0<f_n(x)<\frac{(2n+1)!}{2^{2n}(n!)^2}x^{2n+2}(1-x^2)^{-1/2}. \tag{12.5}$$

From the preceding, we obtain

$$I=\sum_{m=0}^{n}\frac{(2m)!}{2^{2m}(m!)^2}\alpha^{-2m-1}\int_0^\alpha\frac{(1-\cos\beta\cosh v)v^{2m}\,dv}{(\cosh v-\cos\beta)^2}+R_n^{(1)}, \tag{12.6}$$

where

$$|R_n^{(1)}|<\frac{(2n+1)!}{2^{2n}(n!)^2}\alpha^{-2n-3}\int_0^\alpha\frac{|1-\cos\beta\cosh v|v^{2n+2}\,dv}{(\cosh v-\cos\beta)^2\sqrt{1-\left(\frac{v}{\alpha}\right)^2}} \tag{12.7}$$

for $n\geqq 0$. Referring to (9.15), (9.19), and (9.21), we find

$$\begin{aligned}\int_0^{\alpha/2}\frac{|1-\cos\beta\cosh v|v^{2n+2}dv}{(\cosh v-\cos\beta)^2\sqrt{1-\left(\frac{v}{\alpha}\right)^2}}&<\frac{2}{\sqrt{3}}\int_0^{\alpha/2}\frac{|1-\cos\beta\cosh v|}{(\cosh v-\cos\beta)^2}v^{2n+2}\,dv\\ &<\frac{2}{\sqrt{3}}\int_0^{\log 5}(u^{2n}-7.04u^{2n+2}e^{-u})\,du\\ &\quad+\frac{14.08}{\sqrt{3}}\int_0^{\alpha/2}u^{2n+2}e^{-u}\,du\\ &<\frac{14.08}{\sqrt{3}}(2n+2)!\end{aligned} \tag{12.8}$$

for $0\leqq\beta\leqq\pi$ and $\alpha>4n+4$, $n\geqq 0$. We have

$$\int_0^1\frac{u^{2n+2}\,du}{\sqrt{1-u^2}}=\frac{(2n+2)!}{2^{2n+2}[(n+1)!]^2}\left(\frac{\pi}{2}\right)<\frac{1}{2}\sqrt{\frac{\pi}{n+1}}, \tag{12.9}$$

$n\geqq 0$. The inequality follows by use of the asymptotic expansion for the gamma function. As in (9.15), it follows that

$$\frac{|1-\cos\beta\cosh v|}{(\cosh v-\cos\beta)^2}<2.005e^{-v} \tag{12.10}$$

when $v>8$. Using (12.10) and (9.30), and noting that $\alpha^{2n+3}e^{-\alpha/2}$ is a decreasing

function for $\alpha > 4n+6$, we obtain

$$\int_{\alpha/2}^{\alpha} \frac{|1-\cos\beta\cosh v|v^{2n+2}\,dv}{(\cosh v-\cos\beta)^2\sqrt{1-\left(\frac{v}{\alpha}\right)^2}} < 2.005\alpha^{2n+3}\,e^{-\alpha/2}\int_0^1 \frac{u^{2n+2}\,du}{\sqrt{1-u^2}}$$

$$\begin{aligned} &< \frac{2.005(2n+2)!\,\alpha^{2n+3}\,e^{-\alpha/2}}{(2n+2)^{2n+2}\,e^{-2n-2}\sqrt{2\pi(2n+2)}}\sqrt{\frac{\pi}{4n+4}} \\ &= \frac{2.005}{2}(2n+2)!\left(\frac{\alpha}{2n+2}\right)^{2n+3} e^{-\alpha/2+2n+2} \\ &< \frac{2.005}{2}8^3\,e^{-6}(2n+2)! \\ &< 1.273(2n+2)! \end{aligned} \tag{12.11}$$

when $\alpha > 16n+16$, $n \geqq 0$. Hence, it follows that (12.6) is an asymptotic expansion for I, where

$$|R_n^{(1)}| < 9.403\frac{(2n+1)!}{2^{2n}(n!)^2}(2n+2)!\,\alpha^{-2n-3}. \tag{12.12}$$

As in the case of the asymptotic expansion of § 9, (12.6) is also a convergent expansion. In order to obtain a simpler asymptotic expansion, for which the coefficients can be expressed in terms of known functions, we write

$$\begin{aligned} &\sum_{m=0}^{n}\frac{(2m)!}{2^{2m}(m!)^2}\alpha^{-2m-1}\int_0^{\alpha}\frac{(1-\cos\beta\cosh v)v^{2m}\,dv}{(\cosh v-\cos\beta)^2} \\ &\quad = \sum_{m=0}^{n}\frac{(2m)!}{2^{2m}(m!)^2}\alpha^{-2m-1}\int_0^{\infty}\frac{(1-\cos\beta\cosh v)v^{2m}\,dv}{(\cosh v-\cos\beta)^2} + R_n^{(2)}, \end{aligned} \tag{12.13}$$

where

$$R_n^{(2)} = -\sum_{m=0}^{n}\frac{(2m)!}{2^{2m}(m!)^2}\alpha^{-2m-1}\int_{\alpha}^{\infty}\frac{(1-\cos\beta\cosh v)v^{2m}\,dv}{(\cosh v-\cos\beta)^2}. \tag{12.14}$$

It follows from (9.16) and (9.20) that

$$|R_n^{(2)}| < 4.0002\alpha^{-1}\,e^{-\alpha}\sum_{m=0}^{n}\frac{(2m)!}{2^{2m}(m!)^2} \tag{12.15}$$

for $\alpha > 4n+12$. Finally, referring to (9.27), we obtain the asymptotic expansion

$$\lim_{\varepsilon\to 0} D_{12} = \frac{r_2^2\csc^3(\omega_1/2)}{4\pi r_1(r_1+r_2)^2}\left\{\sum_{m=0}^{n}\frac{(-1)^m(2m)!}{(m!)^2}B_{2m}\left(\frac{\beta}{2\pi}\right)\left(\frac{\pi}{\alpha}\right)^{2m+1} + R_n\right\}, \tag{12.16}$$

where

$$R_n = R_n^{(1)} + R_n^{(2)}. \tag{12.17}$$

The preceding estimates for $R_n^{(1)}$ and $R_n^{(2)}$ hold for $\alpha > 16n+16$ and for $0 \leqq \beta \leqq \pi$, $n \geqq 0$.

From (12.12) and (12.9), we obtain

$$\begin{aligned} |R_n^{(1)}| &< 18.806\alpha^{-3}, \quad n=0 \\ &< \frac{9.403(2n+1)}{\sqrt{\pi n}}(2n+2)!\,\alpha^{-2n-3}, \qquad n \geqq 1 \end{aligned} \tag{12.18}$$

when $\alpha > 16n+16$. Proceeding as in (9.30)-(9.36), we find that

$$\alpha^{2n+2} e^{-\alpha} \sum_{m=0}^{n} \frac{(2m)!}{2^{2m}(m!)^2} < \frac{8^2 e^{-14}}{2\sqrt{\pi}} < 0.000016 \tag{12.19}$$

when $\alpha > 16n+16$. From (12.15) and (12.19), it follows that

$$\begin{aligned} |R_n^{(2)}| &< 4.0002\alpha^{-2n-3}\left\{\alpha^{2n+2} e^{-\alpha} \sum_{m=0}^{n} \frac{(2m)!}{2^{2m}(m!)^2}\right\} \\ &< 0.00007(2n+2)!\,\alpha^{-2n-3}. \end{aligned} \tag{12.20}$$

Finally, we obtain

$$\begin{aligned} |R_n| &< 18.9\alpha^{-3}, \quad n=0 \\ &< \frac{9.45(2n+1)}{\sqrt{\pi n}}(2n+2)!\,\alpha^{-2n-3}, \quad n \geqq 1 \end{aligned} \tag{12.21}$$

when $\alpha > 16n+16$ and $0 \leqq \beta \leqq \pi$.

The case $\beta = \pi$ corresponds to the limit $r_2 \to \infty$ with r_1 held constant. Hence, sphere 2 becomes the plane $x=0$ in this limit, and from (10.15), we find

$$\alpha = \pi \cot(\omega_1/2). \tag{12.22}$$

As $\omega_1 \to 0$, we have

$$\begin{aligned} \lim_{\varepsilon \to 0} D_{12} &= \frac{r_2}{r_1(r_1+r_2)} \cdot \frac{1}{\pi\omega_1^2} \cdot [1+O(\omega_1^2)], \qquad 0 < \beta < \pi \\ &= \frac{1}{\pi r_1 \omega_1^2}[1+O(\omega_1^2)], \qquad \beta = \pi. \end{aligned} \tag{12.23}$$

Similarly, the case $\beta = 0$ is obtained by passing to the limit $r_1 \to \infty$ with r_2 held constant. In this limit, sphere 1 becomes the plane $x=0$. Let $(0, y_1, z_1)$ be a given point on the plane $x=0$, with $\rho_1 = \sqrt{y_1^2+z_1^2}$. Then taking the limits $r_1 \to \infty$ and $\omega_1 \to 0$ such that $r_1\omega_1 = \rho_1$, we obtain

$$\begin{aligned} \alpha &= \lim_{r_1 \to \infty} \frac{\pi r_2}{r_1+r_2} \cot\frac{\rho_1}{2r_1} \\ &= \frac{2\pi r_2}{\rho_1}, \end{aligned} \tag{12.24}$$

and

$$\lim_{\varepsilon \to 0} D_{12} = \frac{2r_2^2}{\pi\rho_1^3}\left\{\sum_{m=0}^{n} \frac{(-1)^m(2m)!\,B_{2m}}{(m!)^2}\left(\frac{\pi}{\alpha}\right)^{2m+1} + R_n\right\}, \tag{12.25}$$

where R_n is bounded by (12.21).

13. Expansions for the limiting charge density in terms of inverse radicals. By the procedure of [23, p. 134], we obtain

$$\frac{1}{\cosh z-\cos\beta}=\frac{1}{1-\cos\beta}+\frac{2}{\sin\beta}\left\{\sum_{n=0}^{\infty}\left[\frac{2\pi n+\beta}{z^2+(2\pi n+\beta)^2}-\frac{1}{2\pi n+\beta}\right]\right.$$
$$\left.-\sum_{n=1}^{\infty}\left[\frac{2\pi n-\beta}{z^2+(2\pi n-\beta)^2}-\frac{1}{2\pi n-\beta}\right]\right\} \tag{13.1}$$

for $0<\beta<\pi$ and for all values of z not equal to $\pm i(2\pi n+\beta)$ or $\pm i(2\pi n-\beta)$. We then have

$$\frac{\sinh z}{(\cosh z-\cos\beta)^2}=-\frac{\partial}{\partial z}\frac{1}{\cosh z-\cos\beta}$$
$$=\frac{4z}{\sin\beta}\left\{\sum_{n=0}^{\infty}\frac{2\pi n+\beta}{[z^2+(2\pi n+\beta)^2]^2}-\sum_{n=1}^{\infty}\frac{2\pi n-\beta}{[z^2+(2\pi n-\beta)^2]^2}\right\} \tag{13.2}$$

and

$$\frac{1-\cos\beta\cosh z}{(\cosh z-\cos\beta)^2}=-\frac{\partial}{\partial\beta}\cdot\frac{\sin\beta}{\cosh z-\cos\beta}$$
$$=-2\left\{\sum_{n=0}^{\infty}\frac{z^2-(2\pi n+\beta)^2}{[z^2+(2\pi n+\beta)^2]^2}-\sum_{n=1}^{\infty}\frac{z^2-(2\pi n-\beta)^2}{[z^2+(2\pi n-\beta)^2]^2}\right\} \tag{13.3}$$

when we use the expansion

$$\csc^2\frac{\beta}{2}=4\sum_{n=-\infty}^{\infty}\frac{1}{(2\pi n+\beta)^2}. \tag{13.4}$$

The latter follows, for example, by differentiating an expansion for $\cot(\beta/2)$ in [3, p. 296]. From (13.2) and (10.17), we find

$$\lim_{\varepsilon\to 0}D_{11}=\frac{\pi r_2^2\csc^3(\omega_1/2)}{4r_1(r_1+r_2)^2}\sum_{n=-\infty}^{\infty}\frac{2\pi n+\beta}{[\alpha^2+(2\pi n+\beta)^2]^{3/2}} \tag{13.5}$$

for $0<\beta<\pi$ and $\alpha\geqq 0$, and from (13.3) and (10.18), we have

$$\lim_{\varepsilon\to 0}D_{12}=\frac{\pi r_2^2\csc^3(\omega_1/2)}{4r_1(r_1+r_2)^2}\left\{\sum_{n=0}^{\infty}\frac{2\pi n+\beta}{[\alpha^2+(2\pi n+\beta)^2]^{3/2}}+\sum_{n=1}^{\infty}\frac{2\pi n-\beta}{[\alpha^2+(2\pi n-\beta)^2]^{3/2}}\right\}. \tag{13.6}$$

The interchanges of integration and summation are justified by [3, pp. 499 and 495], respectively. Similarly, from (13.2) and (10.21), we obtain

$$\lim_{\varepsilon\to 0}D_1^*=-\frac{\pi r_2^2\csc^3(\omega_1/2)}{4r_1(r_1+r_2)^2(x_0^2+\rho_0^2)^{1/2}}\sum_{n=-\infty}^{\infty}\frac{2\pi n+\lambda}{[\kappa^2+(2\pi n+\lambda)^2]^{3/2}} \tag{13.7}$$

when $0<\lambda<\pi$, $\kappa\geqq 0$, and (8.5) is satisfied.

We can express (13.5) and (13.6) in terms of ω_1 by substituting (7.7) and (10.15). Denoting r_2/r_1 by b, we obtain

$$\lim_{\varepsilon\to 0}D_{11}=\frac{b^2}{4\pi r_1}\sum_{n=-\infty}^{\infty}\frac{2n(b+1)+b}{\{n^2(b+1)^2-2n(b+1)[n(b+1)+b]\cos\omega_1+[n(b+1)+b]^2\}^{3/2}} \tag{13.8}$$

and

(13.9)
$$\lim_{\varepsilon\to 0} D_{12}=\frac{b^2}{\pi r_1}\left\{\sum_{n=0}^{\infty}\frac{2n(b+1)+b}{\{n^2(b+1)^2-2n(b+1)[n(b+1)+b]\cos\omega_1+[n(b+1)+b]^2\}^{3/2}}\right.$$
$$\left.+\sum_{n=1}^{\infty}\frac{2n(b+1)-b}{\{n^2(b+1)^2-2n(b+1)[n(b+1)-b]\cos\omega_1+[n(b+1)-b]^2\}^{3/2}}\right\}.$$

When $b=1$, these expansions become

(13.10)
$$\lim_{\varepsilon\to 0} D_{11}=\frac{1}{4\pi r_1}\sum_{n=0}^{\infty}\frac{(-1)^n(2n+1)}{[(n+1)^2-2n(n+1)\cos\omega_1+n^2]^{3/2}},$$

(13.11)
$$\lim_{\varepsilon\to 0} D_{12}=\frac{1}{4\pi r_1}\sum_{n=0}^{\infty}\frac{2n+1}{[(n+1)^2-2n(n+1)\cos\omega_1+n^2]^{3/2}}.$$

The expansion (13.8) was given by Poisson [17, pp. 74–79] for $b=1, 2$, and 4 and for $r_1=1$, with the case $b=1$ expressed in the form (13.10).

14. Power series expansions for the limiting charge density. We can express (13.5) and (13.6) as power series in α by expanding each term in powers of α and interchanging the order of summation. The interchange is valid because of the absolute convergence of the double series obtained. We find

(14.1)
$$\lim_{\varepsilon\to 0} D_{11}=\frac{r_2^2\csc^3(\omega_1/2)}{16\pi r_1(r_1+r_2)^2}\cdot\sum_{n=0}^{\infty}(-1)^n\frac{(2n+1)!}{2^{2n}(n!)^2}\left[\zeta\left(2n+2,\frac{\beta}{2\pi}\right)-\zeta\left(2n+2,1-\frac{\beta}{2\pi}\right)\right]\left(\frac{\alpha}{2\pi}\right)^{2n}$$

and

(14.2)
$$\lim_{\varepsilon\to 0} D_{12}=\frac{r_2^2\csc^3(\omega_1/2)}{16\pi r_1(r_1+r_2)^2}\cdot\sum_{n=0}^{\infty}(-1)^n\frac{(2n+1)!}{2^{2n}(n!)^2}\left[\zeta\left(2n+2,\frac{\beta}{2\pi}\right)+\zeta\left(2n+2,1-\frac{\beta}{2\pi}\right)\right]\left(\frac{\alpha}{2\pi}\right)^{2n},$$

where $\zeta(s, a)$ is the generalized zeta function. It follows by the ratio test that (14.1) and (14.2) converge for $|\alpha|<\beta$ when $0<\beta\leqq\pi$. Similarly, from (13.7), we have

(14.3)
$$\lim_{\varepsilon\to 0} D_1^*=-\frac{r_2^2\csc^3(\omega_1/2)}{16\pi r_1(r_1+r_2)^2(x_0^2+\rho_0^2)^{1/2}}\cdot\sum_{n=0}^{\infty}(-1)^n\frac{(2n+1)!}{2^{2n}(n!)^2}\left[\zeta\left(2n+2,\frac{\lambda}{2\pi}\right)-\zeta\left(2n+2,1-\frac{\lambda}{2\pi}\right)\right]\left(\frac{\kappa}{2\pi}\right)^{2n}$$

for $|\kappa|<\lambda$, with $0<\lambda<\pi$.

We see that (14.1) and (14.2) converge for $\pi/2<\omega_1\leqq\pi$, and that α vanishes when $\omega_1=\pi$. It follows from (10.22) that $\kappa=0$ when $\omega_1=2\cot^{-1}h$ and $\phi=\phi_0$, where h is given by (8.14). Hence, (14.3) converges in a region of sphere 1 containing this point in its interior. When $\phi=\phi_0$ and $h\geqq 1$, (14.3) converges for $2\cot^{-1}(h+1)\leqq\omega_1\leqq 2\cot^{-1}(h-1)$. When $h<1$, we find that (14.3) converges for $2\cot^{-1}(h+1)\leqq\omega_1\leqq\pi$ when $\phi=\phi_0$, and for $2\cot^{-1}(1-h)\leqq\omega_1\leqq\pi$ when $\phi=\phi_0+\pi$.

Starting from the expansion (13.4), we can express the coefficients of (14.2) in an alternate form. We have

$$\zeta\left(2n+2,\frac{\beta}{2\pi}\right)+\zeta\left(2n+2,1-\frac{\beta}{2\pi}\right)=(2\pi)^{2n+2}\sum_{n=-\infty}^{\infty}(2\pi n+\beta)^{-2n-2}$$

$$=\frac{(2\pi)^{2n+2}}{4(2n+1)!}\frac{d^{2n}}{d\beta^{2n}}\csc^2\frac{\beta}{2}, \tag{14.4}$$

and hence,

$$\lim_{\varepsilon\to 0} D_{12}=\frac{\pi r_2^2\csc^3(\omega_1/2)}{16r_1(r_1+r_2)^2}\sum_{0}^{\infty}\frac{(-1)^n\alpha^{2n}}{2^{2n}(n!)^2}\left(\frac{d}{d\beta}\right)^{2n}\csc^2\frac{\beta}{2}. \tag{14.5}$$

When $r_1=r_2$, and hence, $\beta=\pi/2$, we have

$$\lim_{\varepsilon\to 0} D_{12}=\frac{\csc^3(\omega_1/2)}{4\pi r_1}\sum_{n=0}^{\infty}(-1)^n\frac{(2n+1)!}{2^{2n}(n!)^2}\left\{\sum_{k=0}^{\infty}(2k+1)^{-2n-2}\right\}\left(\frac{2\alpha}{\pi}\right)^{2n}$$

$$=\frac{\pi\csc^3(\omega_1/2)}{16r_1}\sum_{n=0}^{\infty}\frac{(2^{2n+2}-1)B_{2n+2}}{n!(n+1)!}\alpha^{2n}, \tag{14.6}$$

where B_n is the nth Bernoulli number. We note that we can also obtain (14.2), (14.5), and (14.6) directly by substituting the Maclaurin expansion of $(1-\cos\beta\cosh\alpha u)/(\cosh\alpha u-\cos\beta)^2$ in (10.18) and interchanging integration and summation.

15. Circles of zero charge density. We see that $\lim_{\varepsilon\to 0} D_{11}$ and $\lim_{\varepsilon\to 0} D_{12}$ are finite and positive for $\omega_1=\pi$, and that $\lim_{\varepsilon\to 0} D_{12}$ becomes infinite as $\omega_1\to 0$. When $V_2>V_1>0$, it follows from (3.6) that $\lim_{\varepsilon\to 0} D_1$ is always negative in the neighborhood of $\omega_1=0$, and that it vanishes for some value of ω_1 between zero and π when V_2/V_1 is sufficiently near 1. As V_2/V_1 increases, a value is reached above which the charge on sphere 1 is completely negative.

The power series expansion for the limiting charge density in the neighborhood of $\omega_1=\pi$ is given by

$$\lim_{\varepsilon\to 0} D_{11}=\sum_{n=0}^{\infty}a_n\alpha^{2n},\qquad \lim_{\varepsilon\to 0} D_{12}=\sum_{n=0}^{\infty}b_n\alpha^{2n} \tag{15.1}$$

together with (3.6), where a_n and b_n are defined by (14.1) and (14.2), respectively. When a circle of zero charge density lies sufficiently near the point $\omega_1=\pi$, it satisfies the equation

$$\sum_{n=0}^{\infty}(a_n-\rho b_n)\alpha^{2n}=0, \tag{15.2}$$

where

$$\rho=\frac{V_2-V_1}{V_2+V_1}. \tag{15.3}$$

If $a_1-\rho b_1$ does not vanish when $a_0-\rho b_0=0$, we can invert this series and obtain α^2 as a series in powers of $(a_0-\rho b_0)/(a_1-\rho b_1)$. We see that this condition is satisfied when

$$\frac{b_1}{a_1}<\frac{b_0}{a_0} \tag{15.4}$$

for $0<\beta<\pi$, and we can then verify that $-(a_0-\rho b_0)/(a_1-\rho b_1)>0$ when $a_0-\rho b_0>0$. We see that (15.4) follows from the inequalities

$$1<\frac{\zeta(2,\beta/2\pi)}{\zeta(2,1-\beta/2\pi)}<\frac{\zeta(4,\beta/2\pi)}{\zeta(4,1-\beta/2\pi)} \tag{15.5}$$

for $0<\beta<\pi$. The first inequality is obviously satisfied, since $\beta/2\pi<\frac{1}{2}$, and the second follows from the result that $\zeta(s,a_1)/\zeta(s,a_2)$ is an increasing function of s when $a_2>a_1>0$. To prove the latter result, we use the relation

$$\frac{\zeta(s,a_1)}{\zeta(s,a_2)}=\int_0^\infty \frac{x^{s-1}\,e^{-a_1x}}{1-e^{-x}}\,dx \Big/ \int_0^\infty \frac{x^{s-1}\,e^{-a_2x}}{1-e^{-x}}\,dx \tag{15.6}$$

[23, p. 266], and refer to a theorem by Karlin [8, Thm. 3.4, p. 285]. It follows that (15.4) holds, and hence, that $\alpha^2\sim-(a_0-\rho b_0)/(a_1-\rho b_1)$ as $a_0-\rho b_0\to 0$. Hence, the charge on sphere 1 is completely negative when

$$\frac{V_2}{V_1}>\frac{1+a_0/b_0}{1-a_0/b_0}. \tag{15.7}$$

We find that the right-hand side of (15.7) becomes infinite as $r_1/r_2\to\infty$, and that it tends to 1 as $r_1/r_2\to 0$. Thus, when sphere 2 is a plane, the charge on sphere 1 is completely negative in the limit $\varepsilon\to 0$. When $r_1=r_2$, we have

$$\frac{1+a_0/b_0}{1-a_0/b_0}=\frac{\pi^2+8G}{\pi^2-8G}=6.765\cdots, \tag{15.8}$$

where $G=\sum_{n=0}^{\infty}(-1)^n(2n+1)^2$ is Catalan's constant.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of mathematical functions*, NBS Applied Mathematics Series 55, National Bureau of Standards, Washington, D.C., 1964.

[2] G. K. BATCHELOR AND R. W. O'BRIEN, *Thermal or electrical conduction through a granular material*, Proc. Roy. Soc. London Ser. A, 355 (1977), pp. 313-333.

[3] T. J. I'A. BROMWICH, *An Introduction to the Theory of Infinite Series*, Macmillan, London, 1947.

[4] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. I, McGraw-Hill, New York, 1953.

[5] G. B. JEFFERY, *On a form of the solution of Laplace's equation suitable for problems relating to two spheres*, Proc. Roy. Soc. London Ser. A, 87 (1912), pp. 109-120.

[6] D. J. JEFFREY, *Conduction through a random suspension of spheres*, Proc. Roy. Soc. London Ser. A, 335 (1973), pp. 355-367.

[7] ———, *The temperature field or electric potential around two almost touching spheres*, J. Inst. Math. Appl., 22 (1978), pp. 337-351.

[8] S. KARLIN, *Total Positivity*, Vol. 1, Stanford University Press, Stanford, CA, 1968.

[9] J. B. KELLER, *Conductivity of a medium containing a dense array of perfectly conducting spheres or cylinders or nonconducting cylinders*, J. Appl. Phys., 34 (1963), pp. 991-993.

[10] G. KIRCHHOFF, *Die Vertheilung der Elektricität auf zwei leitenden Kugeln*, J. Reine Angew. Math., 59 (1861), pp. 89-110.

[11] F. KOTTLER, *Electrostatik der Leiter*, Handbuch der Physik, Springer-Verlag, Berlin, 1927, Chap. 4.

[12] J. C. MAXWELL, *A Treatise on Electricity and Magnetism*, Vol. I, Third edition, Oxford University Press, London, 1892.

[13] C. NEUMANN, *Ueber das Gleichgewicht der Wärme und das der Elektricität in einem Körper, welcher von zwei nicht concentrischen Kugelflächen begrenzt wird*, J. Reine Angew. Math., 62 (1863), pp. 36-49.

[14] E. NEUMANN, *Zur Poissonsche Theorie der Elektrostatik, insbesondere über die elektrische Vertheilung auf einem von drei Kugelflächen begrenzten Conductor*, J. Reine Angew. Math., 120 (1899), pp. 60–98, 277–304.

[15] N. E. NÖRLUND, *Vorlesungen Über Differenzenrechnung*, Chelsea, New York, 1954.

[16] G. A. A. PLANA, *Mémoire sur la distribution de l'électricité à la surface de deux sphères conductrices complètement isolées*, Mem. R. Accad. Sci. Torino Ser. 2, 7 (1845), pp. 71–401.

[17] S. D. POISSON, *Mémoire sur la distribution de l'électricité à la surface des corps conducteurs*, Mém. Classe de Sci. Math. Phys. de l'Institut Impériale de France (1811): Part 1, pp. 1–162; Part 2, pp. 163–274.

[18] A. RUSSELL, *The capacity coefficients of spherical electrodes*, Proc. Phys. Soc., 23 (1911), pp. 352–360.

[19] ———, *The capacity coefficients of spherical conductors*, Proc. Roy. Soc. London Ser. A, 97 (1920), pp. 160–172.

[20] ———, *The problem of two electrified spheres*, Proc. Phys. Soc., 35 (1922), pp. 10–29.

[21] ———, *The electrostatic capacity of two spheres when touching one another*, Proc. Phys. Soc., 37 (1925), pp. 282–286.

[22] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Second edition, Cambridge University Press, Cambridge, 1944.

[23] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Fourth edition, Cambridge University Press, Cambridge, 1927; reprinted 1963.

[24] A. VAN TUYL, *The distribution of electricity on two neighboring charged spheres in the presence of an outside point charge*, Ph.D. thesis, Stanford University, Stanford, CA, 1947.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1321-1335, November 1989

003

VARIATIONAL PRINCIPLES FOR EIGENVALUES OF COMPACT OPERATORS*

GILES AUCHMUTY†

Abstract. Variational principles for finding real and complex nonzero eigenvalues, and associated eigenvectors, of a linear compact operator K on a Hilbert space are developed and analyzed. When K is self-adjoint, certain unconstrained variational problems are described for finding the positive, respectively, negative, eigenvalues of K and the corresponding eigenvectors. These principles are extended to generalized eigenproblems and to nonlinear compact operators. For nonself-adjoint linear operators, a minimization problem for certain positive real eigenvalues is described. All the positive real eigenvalues may be described as critical points of a Lagrangian functional. These characterizations are then extended to describe complex eigenvalues and eigenvectors of nonself-adjoint, compact linear operators.

Key words. variational principles, eigenvalue problems, nonself-adjoint compact operators

AMS(MOS) subject classifications. primary 49G05; secondary 47B05

1. Introduction and notation. This paper describes and analyzes some variational principles for the eigenvalues and eigenvectors of a compact operator on a Hilbert space. When the operator is linear, these variational principles are related to those described in [2]–[4] and arise by systematic use of some simple concepts from convex analysis.

Let H be a real, separable Hilbert space with an inner product $\langle\ ,\ \rangle$ and let $K: H \to H$ be a compact linear operator. When K is self-adjoint there is an extensive theory, based on Rayleigh's principle, characterizing the eigenvalues and eigenvectors of K as solutions of various constrained variational problems (see [6] or [10]). In § 2 we describe certain functionals, defined and finite on all of H, whose critical points are eigenvectors of K corresponding to either positive or negative eigenvalues. These functionals have well-defined second derivatives and we can characterize the Morse indices of the critical points. The theory extends related results for self-adjoint matrix eigenproblems as described in § 6 of [3]. Variational principles for generalized eigenproblems are also developed, as are principles for certain eigenvalues and eigenvectors of nonlinear compact operators.

In § 3 these results are extended to obtain a variational principle for some positive eigenvalues, and corresponding eigenvectors, of a nonself-adjoint, compact, linear operator K. The basic constructions here are based on some methods from convex analysis; in particular, the use of conjugate convex functions. The solutions of this variational principle are saddle points of a Lagrangian. We then show, in Theorem 7, that each eigenvector of K, corresponding to a real positive eigenvalue, may be characterized as a critical point of this Lagrangian.

The next section, § 4, generalizes this work and describes how to characterize the complex eigenvalues and eigenvectors of a compact operator as the critical points of a Lagrangian depending on an angular variable θ.

The work described in this paper originated as part of a project to develop variational principles for various nonself-adjoint eigenproblems such as those described in Chandrasekhar's [5] book on hydrodynamic stability. Both theoretically and numerically, there often are advantages to formulating problems as critical point problems.

* Received by the editors April 4, 1988; accepted for publication (in revised form) October 24, 1988. This research was partially supported by National Science Foundation grant DMS 8701886 and by the Air Force Office of Scientific Research.

† Department of Mathematics, University of Houston, Houston, Texas.

Moreover, variational characterizations often lead to good constructive methods for finding solutions. The principles described here appear to be quite different from those obtained for the largest eigenvalues of stochastic matrices as described, for example, in Horn and Johnson [9].

Whenever a term is not defined in this paper, it should be taken in the sense of Zeidler's text [11]. Given a functional $f: H \to \mathbb{R}$, its Gateaux derivative, or gradient, will be denoted by $\nabla f(x)$ and a point $\hat{x}$ in H is a critical point of f whenever $\nabla f(\hat{x}) = 0$. A critical value of f is the value of f at a critical point and the Gateaux second derivative of f will be denoted $D^2 f(x)$.

The functional f is said to be coercive on H provided

$$\liminf_{\|u\| \to \infty} \frac{f(u)}{\|u\|} = \infty.$$

2. Self-adjoint compact operator eigenproblems. Let H be a real Hilbert space and $K: H \to H$ be a compact, linear self-adjoint operator.

Let the positive eigenvalues of K be $\{\lambda_j: j \in J_+\}$. Here J_+ is a subset of the positive integers and we shall assume that

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0 \quad \text{whenever } n \in J_+.$$

Similarly, the negative eigenvalues of K are $\{\lambda_j: j \in J_-\}$ with J_- being a subset of the negative integers and

$$\lambda_{-1} < \lambda_{-2} < \cdots < \lambda_{-n} < 0 \quad \text{whenever } n \in J_-.$$

Either J_+ or J_- may be empty. For each $j \in J_- \cup J_+$, we shall let

$$E_j = \{e \in H: Ke = \lambda_j e \text{ and } \|e\| = 1\} \tag{2.1}$$

be the corresponding set of normalized eigenvectors. When λ_j is a simple eigenvalue of K, then E_j consists of two points. If λ_j is an eigenvalue of multiplicity $m > 1$, then E_j is diffeomorphic to an $(m-1)$-dimensional sphere.

Consider the functional $F_p: H \to \mathbb{R}$ defined by

$$F_p(u) = \frac{1}{p}\|u\|^p - \frac{1}{2}\langle Ku, u\rangle \tag{2.2}$$

with $2 < p < \infty$. The first variational principle we shall analyze is that of minimizing F_p on H. Define

$$\alpha_p = \inf_{u \in H} F_p(u).$$

THEOREM 1. *Suppose H, K, F_p, α_p as above, with $p > 2$; then*

(i) *F_p is weakly lower semicontinuous (w.l.s.c.) and coercive on H and*

$$\alpha_p = \frac{-1}{2r}[\max(0, \lambda_1)]^r \tag{2.3}$$

where $\gamma = (p-2)^{-1}$ and

$$r = p\gamma. \tag{2.4}$$

(ii) *When J_+ is empty, then F_p is minimized at 0; otherwise it is minimized at $\tilde{u} = \lambda_1^{\gamma} e$, where e is in E_1.*

(iii) *The critical points of* F_p *are* 0 *and* $\tilde{u} = \lambda_j^{\gamma} e_j$, *where* j *is in* J_+, *and* e_j *is in* E_j.

(iv) *The critical values of* F_p *are* 0 *and* $(-1/2r)\lambda_j^r$ *for* j *in* J_+.

Proof. Let $f_p(u) = (1/p)\|u\|^p$. Then f_p is continuous and convex; hence it is w.l.s.c. The quadratic form $\langle Ku, u\rangle$ is weakly continuous since K is compact; hence F_p is w.l.s.c. Moreover,

$$F_p(u) \geqq \frac{1}{p}\|u\|^p - \|K\|\,\|u\|^2.$$

Since $p > 2$, this is coercive. Hence F_p is bounded below on H and attains its infimum α_p.

F_p is Gateaux differentiable on H with

$$\nabla F_p(u) = \|u\|^{p-2}u - Ku \tag{2.5}$$

so $\tilde{u}$ is a critical point of F_p if and only if it solves

$$Ku = \|u\|^{p-2}u. \tag{2.6}$$

Hence $\tilde{u} = 0$ or $\tilde{u}$ is an eigenvector of K corresponding to the (positive) eigenvalue $\tilde{\lambda}$ with $\tilde{\lambda} = \|\tilde{u}\|^{p-2}$.

When $\tilde{\lambda} = \lambda_j$, we see that $\tilde{u} = \lambda_j^{\gamma} e_j$ with e_j in E_j.

Take inner products of (2.6) with $\tilde{u}$; then

$$\langle K\tilde{u}, \tilde{u}\rangle = \|\tilde{u}\|^p$$

so

$$F_p(\tilde{u}) = \left(\frac{1}{p} - \frac{1}{2}\right)\|\tilde{u}\|^p = \frac{2-p}{2p}\|\tilde{u}\|^p = \left(\frac{2-p}{2p}\right)\tilde{\lambda}^r. \tag{2.7}$$

Hence (iii) and (iv) follow.

Since $p > 2$, F_p is minimized when $\tilde{\lambda}$ is largest. This occurs at λ_1 when J_+ is nonempty, or at $\tilde{u} = 0$ if J_+ is empty. Thus (2.3), (2.4), and (ii) follow. □

This theorem shows that the unconstrained variational problem of minimizing F_p on H provides a variational principle for finding λ_1 and eigenvectors lying in E_1.

If we can find a $\tilde{u} \in H$ such that $F_p(\tilde{u}) < 0$, it then follows that K has at least one positive eigenvalue λ_1. This can be used to obtain a lower bound on λ_1 as stated in the following result.

COROLLARY 1. *Suppose* $K : H \to H$ *is a compact, linear self-adjoint operator and* F_p *is defined by* (2.2) *with* $p > 2$. *Suppose* $F_p(\tilde{u}) < 0$; *then the largest positive eigenvalue* λ_1 *of* K *obeys*

$$\lambda_1^r = \sup_{u \in C_0} -2rF_p(u) \geqq -2rF_p(\tilde{u}) \tag{2.8}$$

where r *is given by* (2.4) *and* C_0 *is the set of* u *in* H *such that* $F_p(u) < 0$.

Proof. The proof follows from (2.7) and part (ii) of the theorem. □

To find upper bounds on λ_1, one can consider the functional $F_p : H \times [0, \infty) \to \mathbb{R}$ defined by

$$F_p(u, \mu) = \frac{1}{p}\|u\|^p + \frac{\mu}{2}\|u\|^2 - \frac{1}{2}\langle Ku, u\rangle. \tag{2.9}$$

We see that $F_p(u, 0) = F_p(u)$ for all u in H, and for any $\mu_1 > \mu_2 \geqq 0$, we have

$$F_p(u, \mu_1) \geqq F_p(u, \mu_2) \quad \text{for all } u \text{ in } H.$$

Consider the problem of minimizing $F_p(u, \mu)$ on H and let

$$\alpha_p(\mu) = \inf_{u \in H} F_p(u, \mu) \quad \text{with } \mu > 0. \qquad \square$$

COROLLARY 2. *Suppose K, H as above and $F_p(\cdot, \mu)$ is defined by* (2.9) *with $\mu > 0$. Then*

(i) *F_p is w.l.s.c. and coercive on H;*

(ii) (2.10) $$\alpha_p(\mu) = \frac{-1}{2r}[\max(0, \lambda_1 - \mu)]^r$$

is a strictly increasing function of μ on $[0, \lambda_1)$.

Moreover, $\alpha_p(\mu) = 0$ if and only if $\mu \geqq \lambda_1$.

Proof. We have $F_p(u, \mu) = F_p(u) + \mu/2\|u\|^2$.

The extra term here is w.l.s.c. since $\mu > 0$ and helps the coercivity. Thus (i) holds, and

$$\nabla F_p(u, \mu) = \|u\|^{p-2}u + \mu u - Ku.$$

Hence the argument in Theorem 1 provides (2.10). The other results follow from inspection of (2.10).

This provides upper bounds on λ_1, for if we can show that $\inf_{u \in H} F_p(u, \tilde{\mu}) = 0$, then $\lambda_1 \leqq \tilde{\mu}$. $\square$

It is worth noting that we may find the negative eigenvalues of K by replacing K with $-K$ in (2.2). The results may be stated in terms of minimizing

(2.11) $$\tilde{F}_p(u) = \frac{1}{p}\|u\|^p + \frac{1}{2}\langle Ku, u\rangle$$

on H with $p > 2$. Let $\tilde{\alpha}_p = \inf_{u \in H} \tilde{F}_p(u)$.

COROLLARY 3. *Suppose $K : H \to H$ is a self-adjoint, compact linear operator and $\tilde{F}_p$ is defined by* (2.11) *with $p > 2$. Then*

(i) *$\tilde{F}_p$ is w.l.s.c. and coercive on H.*

(ii) *$\tilde{\alpha}_p = -(1/2r)[\max(0, -\lambda_{-1})]^r$ and this is attained at $\tilde{x} = |\lambda_{-1}|^\gamma e$ with e in E_{-1}.*

(iii) *The critical points of $\tilde{F}_p$ are 0 and $|\lambda_j|^\gamma e$ with j in J_- and e in E_j.*

(iv) *The critical values of $\tilde{F}_p$ are 0 and $(-1/2r)|\lambda_j|^r$ for j in J_-.*

Proof. The proof is just like that of Theorem 1, except now

$$\nabla \tilde{F}_p(u) = \|u\|^{p-2}u + Ku$$

and hence the critical points of $\tilde{F}_p$ arise at eigenvectors of K corresponding to negative eigenvalues. $\square$

LEMMA 2.1. *$\nabla F_p : H \to H$ defined by* (2.5) *is Gateaux differentiable with $D^2F_p(0) = -K$ and*

(2.12) $$D^2F_p(u) = \|u\|^{p-2}I - K + (p-2)\|u\|^{p-2}P_u$$

for $u \neq 0$. Here $P_u x = \langle x, u\rangle u/\|u\|^2$ is the projection in the direction u. F_p is convex on H if and only if K is negative semidefinite.

Proof. Consider $\phi(t) = \|u + th\|^{p-2}(u + th)$ for $t \geqq 0$, h in H. We see that

$$t^{-1}(\phi(t) - \phi(0)) = h\|u + th\|^{p-2} + \frac{1}{t}[\|u + th\|^{p-2} - \|u\|^{p-2}]u.$$

Taking limits as t decreases to 0, with $u \neq 0$ we see that

$$\lim_{t \to 0^+} t^{-1}(\phi(t) - \phi(0)) = \|u\|^{p-2}h + (p-2)\|u\|^{p-4}\langle u, h\rangle u.$$

Hence (2.12) follows. When $u = 0$, this limit is zero. F_p will be convex on H if and only if $\langle D^2F_p(u)h, h\rangle \geqq 0$ for all u in H. This implies that $-K$ is positive semidefinite. □

When $\tilde{u}$ is a nonzero critical point of F_p, one may define $\tilde{u}$ to be nondegenerate if zero is not in the spectrum of $D^2F_p(\tilde{u})$. When $\tilde{u}$ is a nonzero, nondegenerate, critical point of F_p, the Morse index of $\tilde{u}$ is defined as the number of negative eigenvalues of $D^2F_p(\tilde{u})$ counting multiplicity. For each j, let d_j be the multiplicity of λ_j as an eigenvalue of K. The following theorem shows that the unconstrained variational problem of extremizing F_p on H has a nice Morse theory. It is a counterpart to the Courant-Fischer-Weyl minimax theory associated with Rayleigh's principle.

THEOREM 2. *Suppose $\tilde{u} = \lambda_j^{\gamma} e$ with e in E_j, j in J_+ being a nonzero critical point of F_p. Then $\tilde{u}$ is a nondegenerate critical point of F_p if and only if λ_j is a simple eigenvalue of K. In this case, the Morse index of $\tilde{u}$ is $\sum_{k=1}^{j-1} d_k$.*

Proof. Let $e^{(k)}$ be a normalized eigenvalue of K that is orthogonal to e. Using the facts that $\|\tilde{u}\|^{p-2} = \lambda_j$, $P_{\tilde{u}} = P_e$, we have

$$D^2F_p(\tilde{u})e^{(k)} = (\lambda_j - \lambda_k)e^{(k)}$$

as $\langle e^{(k)}, e\rangle = 0$. Moreover, $D^2F_p(\tilde{u})e = (p-2)\lambda_j e$.

Since K has a complete orthonormal set of eigenfunctions, so does $D^2F_p(\tilde{u})$. Its eigenvalues are $\lambda_j - \lambda_k$ for $k \neq j$ and $(p-2)\lambda_j$ if λ_j is a simple eigenvalue of K. When λ_j is an eigenvalue of K of multiplicity $d_j > 1$, then we can choose a normalized eigenvector $\tilde{e}$ of K orthogonal to e and such that $D^2F_p(\tilde{u})\tilde{e} = 0$. Hence $\tilde{u}$ will be a degenerate critical point of F_p if and only if λ_j is a simple eigenvalue of K.

When $\tilde{u}$ is nondegenerate, we see from the expressions for $D^2F_p(\tilde{u})e^{(k)}$ that the number of negative eigenvalues of $D^2F_p(\tilde{u})$, counting multiplicity, will be $\sum_{k=1}^{j-1} d_j$. □

It is worth noting that

$$\lim_{p \to \infty} F_p(u) = \chi_1(u) = \begin{cases} 0 & \text{if } \|u\| \leqq 1, \\ \infty & \text{otherwise} \end{cases}$$

where χ_1 is the indicator functional of the unit ball in H.

Rayleigh's principle for finding the largest positive eigenvalue of K is equivalent to minimizing

$$F_\infty(u) = \chi_1(u) - \tfrac{1}{2}\langle Ku, u\rangle.$$

This may be regarded as the convex analysis version of Rayleigh's principle (see [1, § 8]).

The functionals F_p defined by (2.2) define a one-parameter family of variational problems that are smooth, unconstrained and converge, in a certain sense, to Rayleigh's principle as p increases to infinity.

There are a number of interesting and useful extensions of these principles. One may obtain variational principles for other eigenvalues by imposing orthogonality conditions. Suppose one knows the m largest eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_m$ of K and the corresponding eigenspaces. Let $\{e_1, \cdots, e_M\}$ be a corresponding set of orthonormal eigenfunctions of K which is a basis for the direct sum of these eigenspaces. Define

$$H_m = \{u \in H : \langle u, e_j\rangle = 0, 1 \leqq j \leqq M\} \tag{2.13}$$

and consider the problem of minimizing F_p on H_m.

THEOREM 3. *Let $K: H \to H$ be a compact, self-adjoint linear operator and F_p, H_m be defined by* (2.2), (2.13), *respectively, and with $p > 2$. Then*:

(i) *F_p is w.l.s.c. and coercive on H_m.*

(ii) $\alpha_{pm} = \inf_{u \in H_m} F_p(u) = (-1/2r)[\max(0, \lambda_{m+1})]^r$.

(iii) *When $\alpha_{pm} < 0$, this infimum is attained at $\hat{x} = \lambda_{m+1}^{\gamma} e$, where e lies in E_{m+1}.*

Proof. H_m is a closed subspace of H. Since (i) holds on H, it holds on H_m; then α_{pm} is finite and it is attained.

A critical point $\tilde{u}$ of F_p on H_m obeys

$$\nabla F_p(u) = \sum_{j=1}^{M} \mu_j e_j = \|u\|^{p-2} u - Ku \tag{2.14}$$

from the Lagrange multiplier rule. Take inner products of (2.14) with e_k for $1 \leqq k \leqq M$; then, using the self-adjointness of K,

$$\langle K\tilde{u}, e_k \rangle = \lambda_k \langle \tilde{u}, e_k \rangle = \|\tilde{u}\|^{p-2} \langle \tilde{u}, e_k \rangle - \mu_j \delta_{jk}$$

where δ_{jk} is the Kronecker delta.

Since $\tilde{u}$ is in H_m, this implies $\mu_k = 0$ for each $1 \leqq k \leqq M$ and thus $\tilde{u}$ is either zero or an eigenvector of K corresponding to the eigenvalue $\tilde{\lambda} = \|\tilde{u}\|^{p-2}$.

The rest of this proof now parallels that of Theorem 1. □

There are also similar results for finding $\lambda_{-(m+1)}$ by minimizing $\tilde{F}_p$ on the subspace orthogonal to the eigenspaces corresponding to $\lambda_{-1}, \lambda_{-2}, \cdots, \lambda_{-m}$.

Another generalization of these principles is to the generalized eigenproblem of finding nontrivial solutions u of

$$Ku = \lambda Au \tag{2.15}$$

where $A: H \to H$ is a continuous, self-adjoint, linear operator which is also positive definite. That is, there is an $a_0 > 0$ such that

$$\langle Au, u \rangle \geqq a_0 \|u\|^2 \quad \text{for all } u \text{ in } H. \tag{2.16}$$

In this case, define $J_p: H \to \mathbb{R}$ by

$$J_p(u) = \frac{1}{p} \langle Au, u \rangle^{p/2} - \frac{1}{2} \langle Ku, u \rangle \tag{2.17}$$

with $2 < p < \infty$ and consider the problem of minimizing J_p on H.

Let

$$\nu_p = \inf_{u \in H} J_p(u)$$

and $E_{jA} = \{u \in H: Ku = \lambda_j Au \text{ and } \langle Au, u \rangle = 1\}$.

THEOREM 4. *Let $K: H \to H$ be a compact, self-adjoint, linear operator and $A: H \to H$ be continuous, linear, self-adjoint and positive definite. Suppose J_p is defined by* (2.17) *with $p > 2$, then*

(i) *J_p is w.l.s.c. and coercive on H with*

$$\nu_p = \frac{-1}{2r} [\max(0, \lambda_1)]^r$$

and λ_1 is the largest positive eigenvalue of (2.15).

(ii) *When $\nu_p < 0$, then J_p is minimized at $\hat{x} = \lambda_1^{\gamma} e$, where e is in E_{1A}. If $\nu_p = 0$, J_p is minimized at zero.*

(iii) *The critical points of J_p are zero and $\lambda_j^{\gamma} e_j$, where λ_j is a positive eigenvalue of* (2.15) *and e_j is in E_{jA}.*

(iv) *The critical values of J_p are zero and $(-1/2r)\lambda_j^r$, where λ_j is a positive eigenvalue of* (2.15).

Proof. We have that J_p is w.l.s.c. and coercive as before, since A is positive definite. Now

$$\nabla J_p(u) = \langle Au, u\rangle^{(p-2)/2} Au - Ku. \tag{2.18}$$

Hence $\tilde{u}$ is zero or an eigenvector of (2.15) corresponding to the positive eigenvalue

$$\tilde{\lambda} = \langle A\tilde{u}, \tilde{u}\rangle^{(p-2)/2}.$$

Take inner products of (2.18) with $\tilde{u}$; then

$$\langle K\tilde{u}, \tilde{u}\rangle = \langle A\tilde{u}, \tilde{u}\rangle^{(p/2)} = \tilde{\lambda}^r,$$

so

$$J_p(\tilde{u}) = \left(\frac{1}{p} - \frac{1}{2}\right)\tilde{\lambda}^r.$$

Thus $\tilde{u} = \lambda_j^{\gamma} e$, where e is in E_{jA} and λ_j is a positive eigenvalue of (2.15). The proof now parallels that of Theorem 1. □

Finally, it is worth noting that similar variational principles may be studied for nonlinear, compact, potential operators. Let $K: H \to H$ be a nonlinear operator which is the derivative of a functional $\mathcal{K}: H \to \mathbb{R}$ which is weakly continuous and obeys

$$\limsup_{\|u\|\to\infty} \frac{\mathcal{K}(u)}{\|u\|^q} = M \tag{2.19}$$

with M finite and for some $0 \leqq q \leqq \infty$. Then consider $\mathcal{F}: H \to \mathbb{R}$ defined by

$$\mathcal{F}(u) = \frac{1}{p}\|u\|^p - \mathcal{K}(u) \tag{2.20}$$

with $p > \min(1, q)$. We have the following result.

THEOREM 5. *Suppose $\mathcal{K}: H \to \mathbb{R}$ is a weakly continuous, Gateaux-differentiable functional that obeys* (2.19) *and let $\mathcal{F}$ be defined by* (2.20) *with $p > \min(1, q)$. Then*

(i) *F is w.l.s.c. and coercive on H and*

$$\alpha = \inf_{u\in H} \mathcal{F}(u) \text{ is finite.} \tag{2.21}$$

(ii) *The infimum in* (2.21) *is attained at a point $\hat{u}$ in H obeying*

$$\nabla\mathcal{K}(u) = \lambda u \tag{2.22}$$

with

$$\lambda = \|\hat{u}\|^{p-2} \geqq 0. \tag{2.23}$$

Proof. Since $\mathcal{K}$ is weakly continuous, $\mathcal{F}$ is w.l.s.c. Also

$$\frac{\mathcal{F}(u)}{\|u\|^q} = \frac{1}{p}\|u\|^{p-q} - \frac{\mathcal{K}(u)}{\|u\|^q}.$$

Thus

$$\liminf_{\|u\|\to\infty} \frac{\mathcal{F}(u)}{\|u\|^q} \geqq \liminf_{\|u\|\to\infty} \left(\frac{1}{p}\|u\|^{p-q} - M\right) = +\infty$$

as $p > q$, so $\mathcal{F}$ is coercive. Thus α is finite and this infimum is attained. $\mathcal{F}$ is Gateaux differentiable with

$$\nabla\mathcal{F}(u) = \|u\|^{p-2}u - \nabla\mathcal{K}(u)$$

so any minimizer must obey (2.22), (2.23). □

Now consider $\mathcal{F}: H \to \mathbb{R}$ defined by

$$\tilde{\mathcal{F}}(u) = \frac{1}{p}\|u\|^p + \mathcal{K}(u) \tag{2.24}$$

and assume

$$\liminf_{\|u\|\to\infty} \frac{\mathcal{K}(u)}{\|u\|^q} = M > -\infty \tag{2.25}$$

for some $0 \leqq q < \infty$.

COROLLARY 4. *Suppose $\mathcal{K}: H \to \mathbb{R}$ is a weakly continuous, Gateaux-differentiable functional obeying* (2.25) *and that $\tilde{\mathcal{F}}$ is defined by* (2.24) *with $p > \min(1, q)$. Then*

(i) *$\mathcal{F}$ is w.l.s.c. and coercive on H and*

$$\tilde{\alpha} = \inf_{u\in H} \tilde{\mathcal{F}}(u) \text{ is finite.}$$

(ii) *This infimum is attained at a point $\tilde{u}$ in H obeying* (2.22) *with $\tilde{\lambda} = -\|\tilde{u}\|^{p-2} \leqq 0$.*

Proof. The proof is just as before. □

Note that in these results, we have not required any symmetries for $\mathcal{K}$, so there is no reason to believe that $\tilde{u} = 0$ is a critical point of $\mathcal{K}$ or that zero is an eigenvalue of $\nabla\mathcal{K}$ in Theorem 5.

3. Eigenproblems for nonself-adjoint compact operators. Henceforth $K: H \to H$ will be a linear compact operator which is not necessarily self-adjoint. Let $K_s = \frac{1}{2}(K + K^*)$ and $K_a = \frac{1}{2}(K - K^*)$ be the symmetric and antisymmetric parts of K.

Define $F: H \to \mathbb{R}$ by

$$F(u) = \frac{1}{p}\|u\|^p - \frac{1}{2}\langle K_s u, u\rangle \tag{3.1}$$

where $2 < p < \infty$. The conjugate convex function $F^*: H \to \mathbb{R}$ is defined by

$$F^*(v) = \sup_{u\in H} [\langle u, v\rangle - F(u)]. \tag{3.2}$$

We will need the following results about F^*.

LEMMA 3.1. *Suppose F is defined by* (3.1) *with $p > 2$, K_s is a linear, self-adjoint, compact operator and F^* is defined by* (3.2). *Then*

(i) *F^* is w.l.s.c. and convex on H.*

(ii) *$F^*(0) = -\inf_{u\in H} F(u) = (1/2r)[\max(0, \lambda_1)]^r$ where λ_1 is the largest eigenvalue of K_s and r is defined by* (2.4).

(iii) *$F^*(v) \geqq 0$ for all v in H.*

Proof. Since $Q(u, v) = \langle u, v\rangle - F(u)$ is convex and weakly continuous in v for each u in H, this supremum is w.l.s.c. and convex so (i) holds. We have (ii) from Theorem 1, and (iii) holds upon putting $u = 0$ into the right-hand side of (3.2). □

From definition (3.2) of F^* we see that

$$F(u) + F^*(v) \geqq \langle u, v\rangle \tag{3.3}$$

for all (u, v) in $H \times H$. This is sometimes called a generalized Young's inequality. In this problem F need not, in general, be convex, so the strong results of convex analysis do not hold. However, we have the following.

LEMMA 3.2. *Suppose F and F^* are defined by* (3.1) *and* (3.2), *then* (3.3) *holds for all (u, v) in $H \times H$. If equality holds in* (3.3) *at $(\hat{u}, \hat{v})$, then*

$$\nabla F(\hat{u}) = \hat{v}. \tag{3.4}$$

Proof. From (3.2), for each v in H, $F^*(v) \geqq \langle u, v\rangle - F(u)$ for all u in H. Thus (3.3) holds. Moreover, if $F^*(\hat{v}) = \langle \hat{u}, \hat{v}\rangle - F(\hat{u})$, then $\hat{u}$ maximizes $\langle u, v\rangle - F(u)$ on H. Hence it obeys (3.4). □

Define the functional $E : H \to \mathbb{R}$ by

$$E(u) = F(u) + F^*(K_a u) \tag{3.5}$$

and consider the problem of minimizing E on H.

THEOREM 6. *Suppose $K : H \to H$ is a linear compact operator and E, F are as above. Then*

(i) *E is w.l.s.c. and coercive on H, so E attains its infimum on H.*

(ii) *$E(u) \geqq 0$ for all u in H, and if $E(\hat{u}) = 0$ then $\hat{u}$ is a solution of*

$$Ku = \|u\|^{p-2}u. \tag{3.6}$$

Proof. F and F^* are w.l.s.c. from Theorem 1 and Lemma 3.1, respectively. F is coercive from Theorem 1, while F^* is nonnegative. Hence E is coercive and (i) holds.

From (3.3) we have $E(u) = F(u) + F^*(K_a u) \geqq \langle K_a u, u\rangle = 0$ for all u in H and using the antisymmetry of K_a.

From Lemma 3.2, equality holds here provided

$$K_a u = \nabla F(u) = \|u\|^{p-2}u - K_s u,$$

which is (3.6). □

We see that when K is a self-adjoint operator, $K_a = 0$ and then $E(u) = F(u) + F^*(0) = F(u) - \alpha_p$, with α_p being the infimum of F on H.

Note that not every solution of (3.6) obeys $E(\hat{u}) = 0$. In particular we have $E(0) = F^*(0)$ is given by (ii) of Lemma 3.1, and thus when $\lambda_1 > 0$, $E(0)$ will be nonzero.

To illustrate this result, consider the following example.

Let $K : H \to H$ be a compact, linear operator and let $\{e_j : j \in J_1\}$ be an orthonormal set of eigenfunctions of K_s corresponding to nonzero eigenvalues of K_s. Let H_1 be the closed subspace spanned by this set.

Extend this to an orthonormal basis $\{e_j : j \in J\}$ of H and assume $J_2 = J - J_1$ is nonempty. We have

$$\langle K_s u, u\rangle = \sum_{j \in J_1} \lambda_j \langle u, e_j\rangle^2 \tag{3.7}$$

for all u in H, where $\{\lambda_j : j \in J_1\}$ is the set of nonzero eigenvalues of K_s. In this case we can prove the following.

LEMMA 3.3. *Assume K, H as above with F defined by* (3.1) *and* (3.7). *Let λ_1 be the largest positive eigenvalue of K_s and v be orthogonal to H_1. Then*

$$F^*(v) = \begin{cases} \dfrac{1}{q}\|v\|^q & \text{when } \|v\| \geqq \lambda_1^{\nu}, \\ \dfrac{1}{2r}\lambda_1^r + \dfrac{1}{2\lambda_1}\|v\|^2 & \text{when } \|v\| \leqq \lambda_1^{\nu}. \end{cases} \tag{3.8}$$

Here $q = p/p-1$ is the conjugate index to p, $r = p/p-2$ is as in (2.4) *and $\nu = r/q$. When K_s has no positive eigenvalues, then*

$$F^*(v) = \frac{1}{q}\|v\|^q$$

for all v in $H_2 = H \ominus H_1$.

Proof. From (3.2), we see that we have to maximize

$$\langle u, v\rangle - \frac{1}{p}\|u\|^p + \frac{1}{2}\sum_{j\in J_1}\lambda_j\langle u, e_j\rangle^2$$

on H. This functional is weakly upper semicontinuous (w.u.s.c.) and its negative is coercive on H, so it attains its supremum.

The functional is Gateaux differentiable on H, so its supremum must occur at a critical point. Write $u_j = \langle u, e_j\rangle$ for all j in J. The critical points arise at $\hat{u}$, where

$$v_j - \|\hat{u}\|^{p-2}\hat{u}_j + \lambda_j\hat{u}_j = 0 \quad \text{for } j \text{ in } J_1, \tag{3.9}$$

$$v_j - \|\hat{u}\|^{p-2}\hat{u}_j = 0 \quad \text{for } j \text{ in } J_2. \tag{3.10}$$

When v is in H_2, we have $v_j = 0$ for j in J_1 and from (3.10),

$$\|\hat{u}\|^{2(p-2)}\sum_{j\in J_2}\hat{u}_j^2 = \|v\|^2.$$

Let $\hat{u} = \hat{u}^{(1)} + \hat{u}^{(2)}$ be the orthogonal decomposition of $\hat{u}$ defined by the projections onto H_1, H_2. This last equation implies $\|\hat{u}^{(1)}\|^2 = \|\hat{u}\|^2 - (\|v\|/\|\hat{u}\|^{p-2})^2$ upon rearrangement.

If $\hat{u}^{(1)} = 0$ then $\|v\| = \|\hat{u}\|^{p-1}$ and $\langle K_s\hat{u}, \hat{u}\rangle = 0$. Then $F^*(v) = \langle\hat{u}, v\rangle - (1/p)\|\hat{u}\|^p = (1/q)\|v\|^q$.

When $\hat{u}^{(1)} \neq 0$, we must have $\|\hat{u}\|^{p-2} = \lambda_k$ for some k in J_1. In this case

$$\|\hat{u}^{(1)}\|^2 = \lambda_k^{2/p-2} - \frac{\|v\|^2}{\lambda_k^2} = \frac{1}{\lambda_k^2}(\lambda_k^{2\nu} - \|v\|^2)$$

where $\nu = (p-1)/(p-2) = r/q$.

If $\|v\| \geqq \lambda_1^{\nu}$ then this is impossible, so $\hat{u}^{(1)} = 0$. When $\|v\| < \lambda_1^{\nu}$, let $K_1(v) = \{k \in J_1 : \|v\| \leqq \lambda_k^{\nu}\}$. $K_1(v)$ is a finite set when $v \neq 0$, and $\|u^{(1)}\| = \rho_k$, where ρ_k is any of the values $\rho_k^2 = (\lambda_k^{2\nu} - \|v\|^2)/\lambda_k^2$, $k \in K_1(v)$.

Suppose $\|u^{(1)}\| = \rho_k$; then $\|\hat{u}\|^{p-2} = \lambda_k$ and we have

$$\langle\hat{u}, v\rangle - \|\hat{u}\|^p + \langle K_s\hat{u}, \hat{u}\rangle = 0$$

upon taking inner products of (3.9) and (3.10) with $\hat{u}$. Thus the value of (3.9) at $\hat{u}$ is

$$\frac{1}{q}\|\hat{u}\|^p - \frac{1}{2}\langle K_s\hat{u}, \hat{u}\rangle = \frac{1}{q}\lambda_k^r - \frac{1}{2\lambda_k}(\lambda_k^{2\nu} - \|v\|^2).$$

Considered as a function of λ_k, this is maximized when λ_k is largest, that is, at λ_1, and then

$$F^*(v) = \left(\frac{1}{q} - \frac{1}{2}\right)\lambda_1^r + \frac{1}{2\lambda_1}\|v\|^2$$

which yields (3.8) as claimed. □

Suppose now that $K : H \to H$ is a compact linear operator obeying $(K1)$: $H_1 =$ range K_s is a proper closed subspace of H spanned by an orthonormal set $\{e_j : j \in J_1\}$, with $K_s e_j = \lambda_j e_j$ for all j in J_1 and $(K2)$: range $K_a \subseteq H_2 = H_1^{\perp}$.

Then the functional E for this problem is defined by

$$E_p(u) = \frac{1}{p}\|u\|^p - \frac{1}{2}\langle K_s u, u\rangle + F_p^*(K_a u) \tag{3.11}$$

where $F_p^*(v)$ is given by (3.8) and involves only constants r, q depending on p, λ_1 depending on K_s and $\|K_a u\|$.

In particular, we see that $F_p^*(v)$ is an increasing function of $\|v\|$. Thus if K is a compact linear operator obeying $(K1)$ and $(K2)$ then E_p is minimized at $\hat{u}$ in H, where

$\|K_a\hat{u}\| = 0$ and $\hat{u}$ minimizes $F(u) = \frac{1}{2}\|u\|^p - \frac{1}{2}\langle K_s u, u\rangle$. From Theorem 1, this implies $\hat{u} = \lambda_1^\gamma e_1$, where λ_1 is the largest eigenvalue of K_s, e_1 is a corresponding normalized eigenfunction and $\gamma = (p-2)^{-1}$.

If $\tilde{u}$ is a critical point of F corresponding to an eigenvalue λ_j of K_s with $0 < \lambda_j < \lambda_1$, then

$$E_p(\tilde{u}) = \frac{1}{2r}(\lambda_1^r - \lambda_j^r) > 0$$

so $\tilde{u}$ does not minimize E_p on H.

Nevertheless, we have $K_a\tilde{u} = 0$ from assumption $(K2)$ and $K_s\tilde{u} = \lambda_j\tilde{u} = \|\tilde{u}\|^{p-2}\tilde{u}$.

Thus $K\tilde{u} = \|\tilde{u}\|^{p-2}\tilde{u}$, but $E_p(\tilde{u}) > 0$. This shows, again, that the converse of (ii) in Theorem 6 does not hold.

Example. Let Ω be a Lebesgue measurable set in $\mathbb{R}^n$ and $L^2(\Omega)$ be the space of all L^2- integrable, measurable, real-valued functions defined on Ω. This is a Hilbert space with respect to the inner product

$$\langle u, v\rangle = \int_\Omega u(x)v(x)\,dx.$$

Define a linear operator $K: L^2(\Omega) \to L^2(\Omega)$ by

$$Ku(x) = \int_\Omega G(x, y)u(y)\,dy$$

where $G: \Omega \times \Omega \to \mathbb{R}$ is a measurable function obeying $\int_\Omega \int_\Omega |G(x, y)|^2\,dx\,dy < \infty$. Then, from standard results, K is a compact linear operator.

Assume that $G(x, y) = G_1(x, y) + G_2(x, y)$, where

(i) $G_1(x, y) = G_1(y, x)$ a.e. on $\Omega \times \Omega$,

$$G_1(x, y) = \sum_{j=1}^{J} \lambda_j e_j(x)e_j(y)$$

with $\{e_j : 1 \leqq j \leqq J\}$ an orthonormal set in $L^2(\Omega)$, and

(ii) $G_2(x, y) = -G_2(y, x)$ a.e. on $\Omega \times \Omega$ and

$$\int_\Omega G_2(x, y)e_j(y)\,dy = 0 \quad \text{a.e. on } \Omega \text{ for } 1 \leqq j \leqq J.$$

When G obeys (i), (ii), then K obeys conditions $(K1)$, $(K2)$, and we have that K_s is the compact operator defined by the finite-rank kernel G_1 and K_a is the compact operator defined by the kernel G_2.

The functional $E: L^2(\Omega) \to \mathbb{R}$ for this problem is

$$E(u) = \frac{1}{p}\left(\int_\Omega u^2(x)\,dx\right)^{p/2} - \frac{1}{2}\int_\Omega \int_\Omega G_1(x, y)u(x)u(y)\,dx\,dy + F_p^*(K_2u)$$

where $K_2u(x) = \int_\Omega G_2(x, y)u(y)\,dy$ and

$$F_p^*(K_2u) = \begin{cases} \dfrac{1}{q}\left(\displaystyle\int_\Omega |K_2u(x)|^2\,dx\right)^{q/2} & \text{if } \|K_2u\| \geqq \lambda_1^\nu, \\ \dfrac{1}{2r}\lambda_1^r + \dfrac{1}{2\lambda_1}\displaystyle\int_\Omega |K_2u(x)|^2\,dx & \text{if } \|K_2u\| < \lambda_1^\nu. \end{cases}$$

Here $q = p/p-1$, $r = p/p-2$, and we assume $\lambda_1 \geqq \lambda_j$ for all $1 \leqq j \leqq J$, and $\lambda_1 > 0$.

The first two terms in the expression for $E(u)$ constitute the explicit expression for the functional F_p, defined in § 2, when K corresponds to the integral kernel G_1 on $L^2(\Omega)$.

The analysis here depends on the computation of F^* in Lemma 3.3. Other formulae for computing F^* may be based on the results in [7] and [8].

4. Lagrangian formulation. We may find further information about this variational principle for the eigenvalues of nonself-adjoint operators by considering a saddle-point formulation of the problem.

Define $L: H \times H \to \mathbb{R}$ by

$$L(u, v) = \langle K_a u, v\rangle + \frac{1}{p}(\|u\|^p - \|v\|^p) - \frac{1}{2}[\langle K_s u, u\rangle - \langle K_s v, v\rangle] \tag{4.1}$$

$$= \langle K_a u, v\rangle + F_p(u) - F_p(v)$$

where $2 < p < \infty$, and F_p, K_a, K_s are as in § 3.

When L is defined by (4.1), we see that

$$E(u) = \sup_{v \in H} L(u, v) \tag{4.2}$$

upon using (3.2) and (3.5). Thus the primal problem $(\mathcal{P})$ of minimizing E on H is equivalent to finding $\alpha = \inf_{u \in H} \sup_{v \in H} L(u, v)$.

The dual problem is obtained by defining

$$G(v) = \inf_{u \in H} L(u, v). \tag{4.3}$$

It is $(\mathcal{P}^*)$ to maximize G on H and to find

$$\alpha^* = \sup_{v \in H} \inf_{u \in H} L(u, v).$$

The functional L is said to have a saddle point if there is a $(\hat{u}, \hat{v})$ in $H \times H$ such that

$$L(\hat{u}, v) \leqq L(\hat{u}, \hat{v}) \leqq L(u, \hat{v}) \tag{4.4}$$

for all u, v in H. L has a saddle point $(\hat{u}, \hat{v})$ if and only if $\hat{u}$ is a solution of the primal problem $(\mathcal{P})$, $\hat{v}$ is a solution of the dual problem and $\alpha = \alpha^*$. See Theorem 3.1 in [1] for this and related results.

We always have $\alpha^* \leqq \alpha$ and when $\alpha^* < \alpha$ we call $\delta = \alpha - \alpha^*$ the duality gap. The basic properties of the Lagrangian (4.1) may be summarized as follows.

THEOREM 7. *Let $K: H \to H$ be a compact, linear operator and define L by* (4.1) *with $2 < p < \infty$. Then*

(i) *$L(u, v) = -L(v, u)$ for all u, v in H.*

(ii) *$L(., v)$ and $-L(u, .)$ are w.l.s.c. and coercive on H.*

(iii) *Equation* (4.2) *holds, and if G is defined by* (4.3), *then $G(v) = -E(v)$.*

(iv) *L has a saddle point if and only if $\alpha = 0$.*

(v) *L is Gâteaux differentiable on $H \times H$ and $(\hat{u}, \hat{u})$ is a critical point of L if and only if $\hat{u}$ is a solution of* (3.6).

Proof. Part (i) holds by inspection and (ii) holds as in the proof of Theorem 1. Equation (4.2) follows from (3.2)-(3.5) and the last part of (iii) holds from the skew-symmetry of L.

From Theorem 6, we have that α, α^* are finite and are attained, so there exists $\hat{u}$ in H such that $\alpha = E(\hat{u}) = -\alpha^*$.

Thus the duality gap is $\delta = \alpha - \alpha^* = 2\alpha$. Hence when $\alpha \neq 0$, there cannot be a saddle. When $\alpha = 0$, we have from (ii) that there is a $\hat{v}$ in H such that

$$E(\hat{u}) = \sup_{v \in H} L(\hat{u}, v) = L(\hat{u}, \hat{v}). \tag{4.5}$$

But $E(\hat{u}) = 0$ and $L(\hat{u}, \hat{u}) = 0$ so we can take $\hat{v} = \hat{u}$ in (4.5). Then $(\hat{u}, \hat{u})$ will be a saddle point of L when $\alpha = 0$.

L is Gateaux differentiable with respect to both u and v and we have

$$D_u L(u, v) = -K_a v + \|u\|^{p-2} u - K_s u,$$
$$D_v L(u, v) = K_a u - \|v\|^{p-2} v + K_s v.$$

Thus $(\hat{u}, \hat{u})$ will be a critical point of L if and only if $\hat{u}$ obeys (3.6). □

From this theorem, we see that the problem of finding a $\hat{u}$ in H such that $E(\hat{u}) = 0$ is equivalent to finding a saddle point of this Lagrangian. Moreover, the nonzero critical points of this Lagrangian of the form $(\hat{u}, \hat{u})$ are eigenvectors of K corresponding to positive eigenvalues. Thus this Lagrangian provides a critical-point formulation for finding the positive eigenvalues of a general linear compact operator.

We can be more specific about the saddle points of L.

COROLLARY 5. *Under the conditions of Theorem* 7, $(0, 0)$ *is a saddle point of* L *if and only if* $\langle K_s u, u\rangle \leqq 0$ *for all* u *in* H. *Moreover,* $(\hat{u}, \hat{u})$ *is a saddle point of* L *if and only if* $E(\hat{u}) = 0$.

Proof. From (ii) of Lemma 3.1, we have $F^*(0) = 0$ if and only if

$$\langle K_s u, u\rangle \leqq 0 \quad \text{for all } u \in H. \tag{4.6}$$

Thus $E(0) = F^*(0) = 0$ if and only if (4.6) holds. □

From Theorem 7, we see that L has a saddle point if and only if $\alpha = 0$ and in the proof we showed that when $\alpha = 0$ the saddle points have the form $(\hat{u}, \hat{u})$, where $\hat{u}$ is a solution of $E(\hat{u}) = 0$.

The Lagrangian (4.1) is not convex-concave unless (4.6) holds; hence the set of saddle points of L, in general, is not a convex set even when it is nonempty.

This Lagrangian has a well-defined second derivative given by

$$D^2 L(u, v) = \begin{pmatrix} D^2 F(u) & -K_a \\ K_a & -D^2 F(v) \end{pmatrix}$$

with $D^2 F(u)$, $D^2 F(v)$ being defined by the expressions in Lemma 2.1. Just as in Theorem 2, we can use this to define nondegeneracy and type of a critical point of L.

The variational and critical point principles described here appear to be new, even in the finite-dimensional (i.e., matrix) case. In particular, they are different from those described in [4].

5. Complex eigenvalues and eigenvectors. In this section we characterize the complex eigenvalues and eigenvectors of a linear compact operator $K : H \to H$ as the critical points of a real-valued Lagrangian $\mathscr{L}$ depending on an angular variable θ.

Since we want to work only in real arithmetic, we need to rewrite some of the standard definitions. A complex number $\lambda = \lambda_1 + i\lambda_2$ is said to be an eigenvalue of K if there is a nonzero vector (v_1, v_2) in $H \times H$ such that

$$\begin{aligned} Kv_1 &= \lambda_1 v_1 - \lambda_2 v_2, \\ Kv_2 &= \lambda_2 v_1 + \lambda_1 v_2. \end{aligned} \tag{5.1}$$

These are the real and imaginary parts of the usual characterization with $v = v_1 + iv_2$ and $v_1, v_2 \in H$.

If $\lambda = |\lambda| e^{i\theta}$, then this may be rewritten as

$$\mathscr{K}(\theta)\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \cos\theta K & \sin\theta K \\ -\sin\theta K & \cos\theta K \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = |\lambda|\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \tag{5.2}$$

Define $H_2 = H \times H$; then H_2 is a Hilbert space under the inner product $[v, w] = \langle v_1, w_1\rangle + \langle v_2, w_2\rangle$, where $v = \binom{v_1}{v_2}$, $w = \binom{w_1}{w_2}$ are in H_2. Henceforth H_2 will have this inner product, the corresponding norm, and the induced topology. Let $S^1 = [0, 2\pi]$ with the endpoints identified; then for each θ in S^1 we have $\mathcal{K}(\theta): H_2 \to H_2$ is a compact linear operator. From the equivalence of (5.1) and (5.2) we see that K has a nonzero complex eigenvalue $|\lambda|e^{i\theta}$ if and only if $\mathcal{K}(\theta)$ has a positive real eigenvalue $|\lambda|$.

This problem, of characterizing positive real eigenvalues of compact operators, was treated in § 3, so we will extend these results to this case.

Let

$$\mathcal{K}_s(\theta) = \tfrac{1}{2}(\mathcal{K}(\theta) + \mathcal{K}^*(\theta)) = \begin{pmatrix} \cos\theta K_s & \sin\theta K_a \\ -\sin\theta K_a & \cos\theta K_s \end{pmatrix}$$

where K_s, K_a are the symmetric and antisymmetric parts of K. Similarly

$$\mathcal{K}_a(\theta) = \tfrac{1}{2}(\mathcal{K}(\theta) - \mathcal{K}^*(\theta)) = \begin{pmatrix} \cos\theta K_a & \sin\theta K_s \\ -\sin\theta K_s & \cos\theta K_a \end{pmatrix}.$$

Define $G: H_2 \times S^1 \to \mathbb{R}$ by

$$G(v, \theta) = \frac{1}{p}[\|v_1\|^2 + \|v_2\|^2]^{p/2} - \frac{1}{2}[\mathcal{K}_s(\theta)v, v] \tag{5.3}$$

where $v = (v_1, v_2)$ is in H_2 and $2 < p < \infty$.

Then

$$G(v, \theta) = \frac{1}{p}[\|v_1\|^2 + \|v_2\|^2]^{p/2} - \frac{\cos\theta}{2}[\langle K_s v_1, v_1\rangle + \langle K_s v_2, v_2\rangle] + \sin\theta\langle K_a v_1, v_2\rangle$$

in terms of the components v_1, v_2.

Define the conjugate convex function $G^*: H_2 \times S^1 \to \mathbb{R}$ by

$$G^*(w, \theta) = \sup_{v \in H_2} ([v, w] - G(v, \theta))$$

and the functional $\mathcal{G}: H_2 \times S^1 \to \mathbb{R}$ by

$$\mathcal{G}(v, \theta) = G(v, \theta) + G^*(\mathcal{K}_a(\theta)v, \theta). \tag{5.4}$$

Consider the variational principle of minimizing $\mathcal{G}(v, \theta)$ on $H_2 \times S^1$.

THEOREM 8. *Suppose $K: H \to H$ is a compact linear operator and $\mathcal{K}(\theta)$, G, $\mathcal{G}$ are defined as above. Then*

(i) *$\mathcal{G}(., \theta)$ is w.l.s.c. and coercive on H_2 for each θ in S^1;*

(ii) *$\mathcal{G}(v, \theta) \geqq 0$ for all v in H_2;*

(iii) *If $\mathcal{G}(\hat{v}, \hat{\theta}) = 0$ then $\hat{v}$ is a solution of*

$$\mathcal{K}(\hat{\theta})\hat{v} = |\lambda|\hat{v} \tag{5.5}$$

with

$$|\lambda| = \|\hat{v}\|^{p-2}. \tag{5.6}$$

Proof. This follows just as in the proof of Theorem 6. Here we are working with $\mathcal{K}(\theta)$ and H_2 in place of K and H.

The Lagrangian corresponding to this function $\mathcal{G}$ is $\mathcal{L}$: $H_2 \times H_2 \times S^1 \to \mathbb{R}$ defined by

$$\mathcal{L}(v, w, \theta) = \langle \mathcal{K}_a(\theta)v, w\rangle + \mathcal{G}(v, \theta) - \mathcal{G}(w, \theta) \tag{5.7}$$

Again this is skew-symmetric in v and w with

$$\mathscr{L}(w, v, \theta) = -\mathscr{L}(v, w, \theta)$$

and our interest is in finding the critical points of $\mathscr{L}(.,., \theta)$.

THEOREM 9. *Suppose $K: H \to H$ is a compact linear operator and $\mathscr{K}(\theta)$, $\mathscr{L}$ are defined as above. Then $\mathscr{L}(.,., \theta)$ is Gateaux differentiable on H_2 and a nonzero point $(\hat{w}, \hat{w}, \hat{\theta})$ is a critical point of $\mathscr{L}(.,., \hat{\theta})$ if and only if $\hat{w}_1 + i\hat{w}_2$ is an eigenvector of K corresponding to the complex eigenvalue $\hat{\lambda} = \|\hat{w}\|^{p-2} e^{i\hat{\theta}}$.*

Proof. Just as in Theorem 7, we have

$$D_v \mathscr{L}(v, w, \theta) = -\mathscr{K}_a(\theta) w + \|v\|^{p-2} v - \mathscr{K}_s(\theta) v,$$

$$D_w \mathscr{L}(v, w, \theta) = \mathscr{K}_a(\theta) v = \|w\|^{p-2} w + \mathscr{K}_s(\theta) w.$$

Thus $(\hat{w}, \hat{w}, \hat{\theta})$ is a critical point of $\mathscr{L}(.,., \hat{\theta})$ if and only if $\hat{w}$ obeys

$$\mathscr{K}(\hat{\theta}) w = \|w\|^{p-2} w$$

and this is equivalent to having $w_1 + iw_2$ a complex eigenvector of K corresponding to the eigenvalue $\lambda = |\hat{\lambda}| e^{i\hat{\theta}}$ with $|\hat{\lambda}| = \|\hat{w}\|^{p-2}$ as required. □

These last two theorems show that all nonzero eigenvalues and eigenvectors of a compact, linear operator K defined on a real Hilbert space may be found from the critical points of this real-valued functional $\mathscr{L}$. The functional $\mathscr{L}$ is twice continuously Gateaux differentiable and defines an unconstrained problem. Throughout this paper, p can be taken as any number in $(2, \infty)$. The choices $p = 3$ or 4 lead to particularly simple formulae for these linear eigenvalue problems. For nonlinear problems, as described in Theorem 5 and its corollary, the choice of p depends on the properties of K.

REFERENCES

[1] G. AUCHMUTY (1983), *Duality for non-convex variational principles*, J. Differential Equations, 50, pp. 80–145.

[2] ——— (1986), *Dual variational principles for eigenvalue problems*, in Nonlinear Functional Analysis and its Applications, Proc. Symposium in Pure Mathematics, Vol. 45, Part I, American Mathematical Society, Providence, RI, pp. 55–72.

[3] ——— (1989), *Unconstrained variational principles for eigenvalues of real symmetric matrices*, SIAM J. Math. Anal., 20, pp. 1186–1207.

[4] ——— (1989), *Variational principles for eigenvalues of nonsymmetric matrices*, SIAM J. Matrix Anal. Appl., 10, pp. 105–117.

[5] S. CHANDRASEKHAR (1961), *Hydrodynamic and Hydromagnetic Stability*, Oxford University Press, Oxford, London.

[6] J. A. COCHRANE (1972), *The Analysis of Linear Integral Equations*, McGraw-Hill, New York.

[7] R. ELLAIA AND J. B. HIRIART-URRUTY (1986), *The conjugate of the difference of convex functions*, J. Optim. Theory Appl., 49, pp. 493–498.

[8] J. B. HIRIART-URRUTY (1986), *A general formula on the conjugate of the difference of functions*, Canad. Math. Bull., 29, pp. 482–485.

[9] R. A. HORN AND C. A. JOHNSON (1985), *Matrix Analysis*, Cambridge University Press, Cambridge, New York.

[10] H. F. WEINBERGER (1974), *Variational Methods for Eigenvalue Approximation*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia.

[11] E. ZEIDLER (1984), *Nonlinear Functional Analysis and its Application*, III. *Variational Methods and Optimization*, Springer-Verlag, Berlin, New York.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1336-1343, November 1989

004

DECAYING SOLUTIONS OF SEMILINEAR ELLIPTIC EQUATIONS IN $\mathbf{R}^N$*

EZZAT S. NOUSSAIR† AND CHARLES A. SWANSON‡

Abstract. This paper is concerned with the existence and asymptotic behavior of positive solutions of semilinear elliptic problems of second order in $\mathbf{R}^N$, $N \geqq 2$. Positive solutions in $\mathbf{R}^N$ that decay uniformly to zero at ∞ are obtained under various structure conditions by either a direct variational approach or a new approximation procedure. Sharp decay estimates are proved for two general classes of problems.

Key words. elliptic, semilinear, positive solution, asymptotic estimate

AMS(MOS) subject classifications. 35J65, 35P30

1. Introduction. Existence theorems and sharp asymptotic estimates are obtained for positive solutions $u(x)$ in $\mathbf{R}^N$, $N \geqq 2$, of semilinear elliptic eigenvalue problems of the form

$$\begin{aligned} &-\Delta u + b(|x|)u = \lambda p(|x|)f(u), \qquad x \in \mathbf{R}^N, \\ &u \in C^2_{\text{loc}}(\mathbf{R}^N), \qquad \lim_{|x|\to\infty} u(x) = 0 \end{aligned} \tag{1}$$

for some $\lambda > 0$, where b, p are bounded, locally Hölder continuous functions in $\mathbf{R}_+ = [0, \infty)$, $b(r) \geqq 0$, $p(r) > 0$, f is locally Lipschitz continuous in $\mathbf{R}_+$, $f(t) \leqq 0$ for all $t \geqq T$ and some constant T, $f(t) > 0$ for $0 < t < T$, and $f(t) = 0(t^\gamma)$ as $t \to 0+$, $\gamma > 1$. In general, $p(r)$ is not restricted by any monotony or decay conditions at ∞. The conditions on $f(t)$ imply that (1) has a *bounded nonlinearity*, i.e., $f(t)$ is bounded above in $\mathbf{R}_+$.

The extensive literature on the theory and applications of (1), usually dealing with the case of constants b, p or special structure, is indicated in [1], [2], [4]-[9], [11]-[15], and references cited in these sources. Differential equations of this type arise in many scientific areas including quantum field theory, fluid mechanics, astrophysics, gas dynamics, chemistry, and Riemannian geometry.

Our results are separated into two basically distinct cases:

(i) $b(r) \geqq b_0 > 0$; and (ii) $b(r) \geqq 0$. A primary goal is to deal with case (ii) when $p(r)$ does not satisfy a strong decay condition, such as $p(r) = O(r^{-a})$ as $r \to \infty$, $a > 2$. The usual variational approaches in the Sobolev space $W^{1,2}_0(\mathbf{R}^N)$ are not possible in case (ii) since (1) generally has no positive solutions $u \in L^2(\mathbf{R}^N)$, as is well known in the "zero mass" case $b(r) \equiv 0$. Also the method of subsolutions and supersolutions has not been successful in case (ii). The fascination of such problems is in part due to the sensitive dependence of the conclusions on the asymptotic behaviour of $f(u)$ as $u \to 0+$ and $p(r)$ as $r \to \infty$, measured by the constants γ and a in (4) below. To date there is no indication that the difficulties inherent in case (ii) can be handled by the methods of ordinary differential equations, e.g., the shooting method, trajectory analysis, fixed-point theorems, or Lyapunov techniques. Instead, our method employs a sequence of differential equations

$$-\Delta u_k + \left[b(|x|) + \frac{1}{k} \right] u_k = \lambda_k p(|x|) f(u_k), \qquad x \in \mathbf{R}^N, \tag{2}$$

* Received by the editors October 5, 1987; accepted for publication (in revised form) February 8, 1989.
† School of Mathematics, University of New South Wales, Kensington, New South Wales, Australia 2033.
‡ Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Y4. The research of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A-3105.

of type (i), with positive radial solutions $u_k \in W_0^{1,2}(\mathbf{R}^N)$ and corresponding $\lambda_k > 0$, $k = 1, 2, \cdots$. The idea is to prove that $\{u_k\}$ has a subsequence that converges locally uniformly in $C^2(\mathbf{R}^N)$ to a positive solution u of (1) as $k \to \infty$. For the full γ range described by (4), this is the first procedure devised to date that can treat the zero mass case (and generalizations) regardless of the nonlinear structure of $f(u)$.

A necessary prerequisite for this program is the construction of such positive pairs (λ_k, u_k) of (2) without the assumption of any monotony or decay conditions on $p(r)$ as $r \to \infty$. As far as we are aware, such a theorem is not contained among similar results of Berestycki and Lions [1], [2], Gidas, Ni, and Nirenberg [5], Strauss [13], Stuart [14], and others. Accordingly, we have sketched a proof of Theorem 1 below that we believe is of independent interest in view of its relevance to nonlinear field theory [2], [13].

THEOREM 1. *In case* (i), *problem* (1) *has a positive solution pair* (λ, u_λ), *with* $u_\lambda \in W_0^{1,2}(\mathbf{R}^N)$, *such that* $u_\lambda(x)$ *and* $|\nabla u_\lambda(x)|$ *are asymptotically bounded by constant multiples of* $|x|^{(1-N)/2} \exp(-\sqrt{b_0}\,|x|)$ *as* $|x| \to \infty$.

In case (ii) a positive solution pair (λ, u_λ) of (1) does not exist in general without restrictions on p or γ. An example showing this is

$$\begin{aligned} &-\Delta u = \lambda(1+|x|^a)^{-1}(u^\gamma - u^\beta), \qquad x \in \mathbf{R}^N, \quad N \geqq 3, \\ &u \in C^2_{\mathrm{loc}}(\mathbf{R}^N), \qquad \lim_{|x|\to\infty} u(x) = 0 \end{aligned} \tag{3}$$

for $a \geqq 0$, $1 < \gamma < \beta$. In fact, if such a solution pair exists, then u_λ satisfies

$$-\Delta u_\lambda \geqq (\lambda/2)(1+|x|^a)^{-1} u_\lambda^\gamma$$

for all sufficiently large $|x|$, implying $\gamma > (N-a)/(N-2)$ by a well-known oscillation criterion [10, p. 76]. In the case where $a = 2$, it therefore follows from Theorem 2 below that the condition $\gamma > 1$ is necessary and sufficient for the existence of a positive solution pair (λ, u_λ).

For Theorem 2 we impose the additional hypothesis

$$\begin{aligned} &p(r) = O(r^{-a}) \quad \text{as } r = |x| \to \infty, \quad N \geqq 3, \\ &\gamma > 1 \quad \text{if } a \geqq 2, \qquad \gamma > \frac{N-2a+2}{N-2} \quad \text{if } 0 \leqq a < 2. \end{aligned} \tag{4}$$

THEOREM 2. *In case* (ii), *if* (4) *holds, problem* (1) *has a positive solution pair* (λ, u_λ) *with* $u_\lambda \in L^{2N/(N-2)}(\mathbf{R}^N)$ *and* $u_\lambda(r) = O(r^{(2-N)/2})$ *uniformly as* $r \to \infty$.

For a domain $G \subset \mathbf{R}^N$, the norms in $L^q(G)$, $L^q(\mathbf{R}^N)$, and $E = W_0^{1,2}(\mathbf{R}^N)$ will be denoted by $\|\cdot\|_{q,G}$, $\|\cdot\|_q$, and $\|\cdot\|_{1,2}$, respectively. We will use the notation

$$f_T(t) = \begin{cases} f(t) & \text{if } 0 \leqq t \leqq T, \\ 0 & \text{otherwise,} \end{cases}$$

$$F(t) = \int_0^t f(s)\,ds, \qquad F_T(t) = \int_0^t f_T(s)\,ds,$$

$$I(\phi) = \frac{1}{2}\int_{\mathbf{R}^N} [|\nabla\phi|^2 + b\phi^2]\,dx, \qquad \phi \in E,$$

$$J(\phi) = \int_{\mathbf{R}^N} pF_T(\phi)\,dx, \qquad \phi \in E,$$

$$(Pv)(x) = \frac{1}{\mu(S_1)}\int_{S_1} v(|x|\omega)\,d\mu(\omega),$$

where μ denotes the usual surface measure on the unit sphere S_1 in $\mathbf{R}^N$. The operator P is an orthogonal projection from $L^2(\mathbf{R}^N)$ into radial L^2 functions. It is known [14, Lemma 6.1] that P extends to a bounded linear mapping from E into E such that

$$\|\nabla Pv\|_2 \leqq \|P|\nabla v|\,\|_2 \leqq \|\nabla v\|_2, \qquad v \in E. \tag{5}$$

Geometrically $(Pv)(x) = V(r)$ represents the spherical mean of $v(x)$ over the sphere S_r of radius r in $\mathbf{R}^N$.

Proof of Theorem 1. Let $\{v_n\}$ be a weakly convergent sequence in E, with weak limit v, and define $u = Pv$, $u_n = Pv_n$. Then [14, Lemma 6.1] $\{u_n\}$ is uniformly bounded in the norm $\|\cdot\|_{1,2}$ in E, and so also in the norm $\|\cdot\|_2$. By the hypothesis that $p(x)$ is bounded and $f(t) = O(t^\gamma)$ as $t \to 0$, there exists a constant $C > 0$ such that

$$\begin{aligned} |J(u_n) - J(u)| &\leqq \int_{\mathbf{R}^N} \int_{u(x)}^{u_n(x)} C|t|^\gamma \, d|t| \, dx \\ &\leqq 2^{\gamma-1} C \int_0^1 [G(u_n, u, \theta) + H(u_n, u, \theta)] \, d\theta \end{aligned} \tag{6}$$

by Fubini's Theorem in $L^1(\mathbf{R}^N \times [0,1])$ and a slight computation, where

$$G(u_n, u, \theta) = \int_B |u_n(x) - u(x)|[|u(x)|^\gamma + \theta^\gamma |u_n(x) - u(x)|^\gamma] \, dx, \tag{7}$$

$$H(u_n, u, \theta) = \int_{B'} |u_n(x) - u(x)|[|u(x)|^\gamma + \theta^\gamma |u_n(x) - u(x)|^\gamma] \, dx, \tag{8}$$

and B' denotes the complement of the unit ball B in $\mathbf{R}^N$. The Cauchy-Schwarz inequality yields

$$G(u_n, u, \theta) \leqq \|u_n - u\|_{2,B}[\|u\|_{2\gamma,B}^\gamma + \theta^\gamma \|u_n - u\|_{2\gamma,B}^\gamma],$$

and hence the compactness of the Sobolev embedding $W_0^{1,2}(B) \subset L^q(B)$ for $2 \leqq q \leqq 2N/(N-2)$, $N \geqq 3$ (or $q \geqq 1$, $N = 2$) guarantees the existence of a subsequence of $\{u_n\}$, again denoted by $\{u_n\}$, such that $\lim_{n\to\infty} G(u_n, u, \theta) = 0$.

To obtain an analogue of this for (8), we use an a priori estimate of Strauss [13, Lemma 1] for radial functions $z \in E$:

$$|z(r)| \leqq C_0 \zeta(r) \|z\|_{1,2}, \qquad |x| = r \geqq 1 \tag{9}$$

where $\zeta(r) = r^{(1-N)/2}$, $N \geqq 2$, and C_0 is a positive constant independent of z. Application of (9) to (8) shows that there exists a constant K, independent of n, such that

$$H(u_n, u, \theta) \leqq K \|\zeta^{\gamma-1}(u_n - u)\|_{2,B'}[\|u\|_{2,B'} + \theta^\gamma \|u_n - u\|_{2,B'}].$$

From the decay of $\zeta(r)$ as $r \to \infty$, multiplication by $\zeta^{\gamma-1}$ is a compact operator from $W_0^{1,2}(B')$ into $L^2(B')$ by a theorem of Berger and Schechter [3, Thm. 2.7], i.e., $\lim_{n\to\infty} \|\zeta^{\gamma-1}(u_n^* - u)\|_{2,B'} = 0$ for a subsequence $\{u_n^*\}$ of $\{u_n\}$, implying that $\lim_{n\to\infty} H(u_n^*, u, \theta) = 0$. It then follows from (6)-(8), if $v_n \to v$ weakly in E, that

$$\lim_{n\to\infty} J(Pv_n^*) = J(Pv) \tag{10}$$

for a subsequence $\{v_n^*\}$ of $\{v_n\}$, i.e., the functional J_p: $v \to J(Pv)$ is weakly sequentially compact.

By the continuity of p and F_T, there exists a nontrivial radial function $\phi \in E$ such that $d = J(\phi) > 0$. Consider the constrained minimization problem

$$M = \inf\{I(v): v \in E, J(Pv) = d\},$$

and let $\{v_n\}$ be a minimizing sequence in E with $\lim_{n\to\infty} I(v_n) = M$. Since I is an equivalent norm on E, $\{v_n\}$ is bounded in E and therefore has a weakly convergent subsequence, also denoted by $\{v_n\}$, to a function $v \in E$. Since (10) holds for a subsequence $\{v_n^*\}$, it follows that $Pv \not\equiv 0$ since $d = J(Pv)$ and $J(0) = 0$. The Euler-Lagrange principle shows that $I'(v) = \lambda J'(Pv)$ for some real λ, where I' and J' denote Fréchet derivatives. By the assumption $f(t) = o(t)$ as $t \to 0$, standard procedure enables us to write the last functional equation as

$$\int_{\mathbf{R}^N} [\nabla v \cdot \nabla w + bvw]\, dx = \lambda \int_{\mathbf{R}^N} pf_T(Pv) Pw\, dx \tag{11}$$

for all $w \in E$. Since $Pw \in E$ for all $w \in E$, this can be rewritten as [14, Lemma 6.1]

$$\int_{\mathbf{R}^N} [\nabla(Pv) \cdot \nabla w + b(Pv)w]\, dx = \lambda \int_{\mathbf{R}^N} pf_T(Pv) w\, dx \tag{12}$$

for all $w \in E$, from which $u = Pv$ is a nontrivial weak solution of the problem

$$-\Delta u + bu = \lambda p f_T(u), \qquad u \in E. \tag{13}$$

Since (12) holds in particular for $w = Pv$, it follows that $\lambda \neq 0$. In view of the regularity hypotheses on b, p, f, standard elliptic regularity theory shows that $u \in C^{2+\alpha}_{\text{loc}}(\mathbf{R}^N)$ for some $\alpha \in (0, 1)$, i.e., u is a classical solution of (13).

To verify that $u \geqq 0$ throughout $\mathbf{R}^N$, suppose to the contrary that $u < 0$ in a nonempty set Ω. The definition of f_T shows that $-\Delta u + bu = 0$ in Ω and $u|_{\partial\Omega} = 0$, implying the contradiction $u(x) \equiv 0$ in Ω by the maximum principle.

Since $u(x)$ is nontrivial and nonnegative in $\mathbf{R}^N$, (12) with $w = Pv = u$ shows that $\lambda > 0$. Also $0 \leqq u(x) \leqq T$ in $\mathbf{R}^N$, for if $u(x_0) > T$ is a positive maximum of a solution of (13), then in appropriate coordinates

$$\lambda p(|x|) f_T(u(x_0)) = -(\Delta u)(x_0) + b(|x_0|)u(x_0) \geqq b_0 u(x_0) > 0,$$

contrary to the definition of $f_T(t)$. Therefore $f_T(u) = f(u)$ in (13), and hence u is a solution of (1). Since $f(t) \geqq 0$ for $0 \leqq t \leqq T$, $-\Delta u + bu \geqq 0$ in $\mathbf{R}^N$, implying that $u(x) > 0$ throughout $\mathbf{R}^N$ by the strong maximum principle. The decay estimates in Theorem 1 can be deduced from the property $\lim_{|x|\to\infty} u(x) = 0$ uniformly in $\mathbf{R}^N$ (since $u \in E$) by a slight modification of the argument of Gidas, Ni, and Nirenberg [5, Prop. 4.1].

3. Proof of Theorem 2. The idea of the proof is to apply Theorem 1 to the sequence of equations (2) and appeal to the three lemmas below. We will use the following additional notation:

$$I_k(\phi) = \frac{1}{2} \int_{\mathbf{R}^N} \left[|\nabla \phi|^2 + \left(b + \frac{1}{k} \right) \phi^2 \right] dx, \quad \phi \in E, \quad k = 1, 2, \cdots. \tag{14}$$

LEMMA 1. *Under the assumptions of Theorem* 2, *for every* $k = 1, 2, \cdots$ *there exists a positive solution pair* (λ_k, u_k) *of* (2) *such that* $u_k \in E$, *the sequence* $\{\|\nabla u_k\|_2\}$ *is bounded, and*

$$0 < u_k(r) \leqq M r^{(2-N)/2}, \qquad r \geqq 1 \tag{15}$$

for a positive constant M *independent of* k.

Proof. Such a sequence (λ_k, u_k) is guaranteed by Theorem 1, where $u_k = Pv_k$ and v_k is a solution of the constrained minimization problem

$$I_k(v_k) = \inf\{I_k(v): v \in E, J(Pv) = d\}. \tag{16}$$

It follows from (5), (14), and (16) that

$$\|\nabla u_k\|_2^2 = \|\nabla Pv_k\|_2^2 \leqq \|\nabla v_k\|_2^2 \leqq 2I_k(v_k) \leqq 2I_1(v_1).$$

Hence $\|\nabla u_k\|_2$ is bounded, implying (15) by the estimate of Berestycki and Lions for radial functions in E [2, Lemma AIII].

LEMMA 2. *The sequence $\{u_k\}$ in Lemma 1 has a subsequence that converges locally uniformly in $C^2(\mathbf{R}^N)$ to a nontrivial function $u \in C^2_{\text{loc}}(\mathbf{R}^N)$ satisfying*

$$0 \leqq u(r) \leqq Mr^{(2-N)/2}, \qquad r \geqq 1. \tag{17}$$

Proof. The existence of such a convergent subsequence is a consequence of the uniform estimate (15) by a standard argument based on Schauder and L^q-estimates in bounded domains of $\mathbf{R}^N$ (see, e.g., [11, Thm. 4.3]). To prove that the limit $u(r)$ is not identically zero, we use (15), the uniform bound $0 < u_k(r) \leqq T$ in $\mathbf{R}^N$, and the assumption $F(t) = O(t^{\gamma+1})$ as $t \to 0+$ to conclude that there exists a positive constant C, independent of k, such that

$$\begin{aligned} 0 < d = J(u_k) &= \int_{\mathbf{R}^N} p(|x|)F(u_k(|x|))\,dx \\ &\leqq C \int_{\mathbf{R}^N} p(|x|)[u_k(|x|)]^{\gamma+1}\,dx. \end{aligned} \tag{18}$$

By (4) and (15), $p(r)[u_k(r)]^{\gamma+1} \leqq C_1 r^{-b}$ for $r \geqq 1$, where $C_1 > 0$ is another constant independent of k and

$$b = a + \frac{(N-2)(\gamma+1)}{2} > N$$

since $(N-2)(\gamma+1) > 2N-2a$ by (4). Therefore $pu_k^{\gamma+1} \in L^1(\mathbf{R}^N)$, and since $\{pu_k^{\gamma+1}\}$ converges pointwise to $pu^{\gamma+1} \in L^1(\mathbf{R}^N)$, the Dominated Convergence Theorem applied to (18) shows that

$$0 < d \leqq C \int_{\mathbf{R}^N} p(|x|)[u(|x|)]^{\gamma+1}\,dx,$$

proving that u is nontrivial.

LEMMA 3. *The sequence $\{\lambda_k\}$ in Lemma 1 is bounded.*

Proof. The analogue of (11) for (λ_k, u_k) is (taking $w = v = v_k$, $u_k = Pv_k$ in (11))

$$\int_{\mathbf{R}^N} \left[|\nabla v_k|^2 + \left(b + \frac{1}{k}\right) v_k^2\right] dx = \lambda_k \int_{\mathbf{R}^N} pu_k f(u_k)\,dx,$$

implying by (14) and (16) that

$$2I_1(v_1) \geqq 2I_k(v_k) = \lambda_k \int_{\mathbf{R}^N} pu_k f(u_k)\,dx, \tag{19}$$

$k = 1, 2, \cdots$. It follows as in Lemma 2 that $pu_k f(u_k) \in L^1(\mathbf{R}^N)$ and $\{pu_k f(u_k)\}$ converge pointwise to $puf(u) \in L^1(\mathbf{R}^N)$. By Fatou's Lemma,

$$\liminf_{k\to\infty} \int_{\mathbf{R}^N} pu_k f(u_k)\,dx \geqq \int_{\mathbf{R}^N} puf(u)\,dx = \delta > 0$$

since u is not identically zero by Lemma 2 and $f(u)>0$ in $0<u<T$. It is then a consequence of (19) that $2I_1(v_1)\geqq\delta\lambda_k$, and hence $\{\lambda_k\}$ is bounded.

To complete the proof of Theorem 2, let λ be the limit of a convergent subsequence of $\{\lambda_k\}$. Since a subsequence of $\{u_k\}$ converges to u in $C^2_{\text{loc}}(\mathbf{R}^N)$ we can let $k\to\infty$ in (2) to conclude that $u=u_\lambda$ is a solution of the differential equation in (1). Since $\|\nabla u_k\|_2$ is bounded by Lemma 1, $\{\|u_k\|_{2N/(N-2)}\}$ also is bounded by a standard Sobolev Embedding Theorem. In view of the convergence of $\{u_k\}$ to u_λ, Fatou's Lemma shows that $u_\lambda\in L^{2N/(N-2)}(\mathbf{R}^N)$. We can prove that $|(\nabla u)(|x|)|$, as well as $u(|x|)$, has uniform limit zero as $|x|\to\infty$ by use of interior Hölder estimates for (1).

Since $\lambda\geqq 0$, it must be that $\lambda>0$, for otherwise u_λ would be a nontrivial, nonnegative solution of $\Delta u_\lambda=0$ in $\mathbf{R}^N$ with uniform limit zero at ∞, contradicting the maximum principle. The strict positivity of u_λ throughout $\mathbf{R}^N$ then follows from the strong maximum principle.

4. Asymptotic estimates. The asymptotic estimate (17) for the solution $u(r)$ of (1) in Theorem 2 can be sharpened to

$$u(r)=O(r^{2-N+\varepsilon})\quad\text{as } r=|x|\to\infty \tag{20}$$

for arbitrary $\varepsilon>0$. To prove (20), first note from (4) that

$$\gamma-1=2(2-a+2\delta)/(N-2) \tag{21}$$

for some $\delta>0$. Consider the recursive sequence defined by

$$\sigma_0=\frac{N-2}{2},\quad \sigma_{k+1}=\frac{1}{2}[(a-2)+\sigma_k(\gamma+1)],\quad k=0,1,2,\cdots. \tag{22}$$

Then $\sigma_k\geqq\sigma_0$ by induction and hence

$$\sigma_{k+1}-\sigma_k=\tfrac{1}{2}[a-2+\sigma_k(\gamma-1)]\geqq\delta,\qquad k=0,1,2,\cdots,$$

implying that $\{\sigma_k\}$ is increasing and $\sigma_k\to\infty$ as $k\to\infty$. Let m be the smallest integer for which $\sigma_{m+1}\geqq N-2-\varepsilon$. We know from (17) that $u(r)=O(r^{-\sigma_0})$ as $r\to\infty$. If $u(r)=O(r^{-\sigma_{k-1}})$ for an integer k, $1\leqq k\leqq m$, consider the function $v_A(r)=A(1+r^2)^{-\sigma_k/2}$, $r\geqq 0$, for a constant $A>0$. A calculation gives

$$-(\Delta v_A)(r)=\frac{A\sigma_k[(N-2-\sigma_k)r^2+N]}{(1+r^2)^{(\sigma_k+4)/2}},$$

whereas the assumptions on p and f imply that $u(r)$ satisfies

$$-(\Delta u)(r)\leqq\lambda p(r)f(u(r))\leqq Cr^{-a-\gamma\sigma_{k-1}},\qquad r\geqq 1$$

for some positive constant C. Since $\sigma_k<N-2-\varepsilon$ for $k\leqq m$, and $\sigma_k+2<a+\gamma\sigma_{k-1}$ by (21) and (22), positive constants A and R can be selected such that

$$-(\Delta v_A)(r)\geqq-(\Delta u)(r)\quad\text{for } r\geqq R,$$
$$v_A(R)\geqq u(R).$$

Since $v_A(r)\to 0$ and $u(r)\to 0$ as $r\to\infty$, the maximum principle implies that $v_A(r)\geqq u(r)$ for all $r\geqq R$, and hence $u(r)=O(r^{-\sigma_k})$ as $r\to\infty$. In particular, $u(r)=O(r^{-\sigma_m})$ as $r\to\infty$.

To complete the proof of (20) we redefine $\sigma'_{m+1}=N-2-\varepsilon$, so $\sigma'_{m+1}\leqq\sigma_{m+1}$ by our choice of m. The argument above can then be repeated with σ_{k-1}, σ_k replaced by σ_m, σ'_{m+1}, respectively, in view of the inequalities

$$\sigma'_{m+1}+2\leqq\sigma_{m+1}+2<a+\gamma\sigma_m,$$

resulting in $u(r)=O(r^{-\sigma'_{m+1}})$ as $r\to\infty$.

Under any one of the three additional hypotheses I, II, or III below, ε can be replaced by zero in the asymptotic estimate (20), i.e., the solution $u(r)$ of (1) in Theorem 2 has a "removable singularity" at ∞:

(I) $\liminf_{r\to\infty} r^{\nu}b(r) = +\infty$ for some $\nu < a+(N-2)(\gamma-1)$.

(II) $\gamma > (N-a+2)/(N-2)$.

(III) $b(r) \equiv 0$, f, $p \in C^1(0,\infty)$, and there exist positive constants C_1, C_2, C_3, C_4, t_0, and r_0 such that

(A) $C_1 t^{\gamma} \leqq f(t) \leqq C_2 t^{\gamma}$ for $0 < t \leqq t_0$, where γ satisfies (4) for $0 \leqq a < 2$;

(B) $C_3 r^{-a} \leqq p(r) \leqq C_4 r^{-a}$ for $r \geqq r_0$;

(C) $(\log p)'(r) = O(r^{-1})$ as $r \to \infty$, $(\log g)'(t) = O(1)$ as $t \to 0+$, where $g(t) = t^{-\gamma} f(t)$, $t > 0$.

The *upper* estimates for $f(t)$ and $p(r)$ in (A) and (B) have already been imposed, of course, as hypotheses for Theorem 2. The proof below in case (III) is based on an a priori estimate of Gidas and Spruck [6, Thm. 3.6(iii)] for equations of type $-\Delta u = h(r)u^{\gamma}$, requiring extra conditions similar to (B) and (C) above.

COROLLARY. *If any one of the conditions* (I)-(III) *holds, the solution* $u(r)$ *of* (1) *in Theorem* 2 *satisfies* $u(r) = O(r^{2-N})$ *and* $u'(r) = O(r^{1-N})$ *as* $r \to \infty$.

Proof. (I) Since $f(t) = O(t^{\gamma})$ as $t \to 0+$, it follows from (1), (20), for a sufficiently large constant $R_0 > 0$, that

$$\begin{aligned} -\Delta u &\leqq [Cr^{-a}u^{\gamma-1} - b(r)]u \\ &\leqq [\tilde{C}r^{-\nu} - b(r)]u \leqq 0 \end{aligned} \tag{23}$$

for all $r \geqq R_0$, where C, $\tilde{C}$ are positive constants and

$$\nu = a+(N-2-\varepsilon)(\gamma-1) < a+(N-2)(\gamma-1).$$

The comparison function $v(r) = Ar^{2-N}$ satisfies $\Delta v = 0$ and $v(R_0) \geqq u(R_0)$ for a sufficiently large constant $A > 0$. Let $w = v - u$. Then

$$-\Delta w \geqq 0 \quad \text{for } r = |x| \geqq R_0, \ w(R_0) \geqq 0, \quad \text{and} \quad w(r) \to 0 \quad \text{as } r \to \infty,$$

implying that $u(r) \leqq v(r)$ for all $r \geqq R_0$ by the maximum principle.

(II) As in (23), u satisfies

$$-\Delta u \leqq Cr^{-a-\gamma(N-2-\varepsilon)} \quad \text{for } r \geqq R_0.$$

A calculation shows that $v(r) = A(1+r^2)^{(2-N)/2}$ is a solution of

$$-\Delta v = AN(N-2)(1+r^2)^{-(N+2)/2}, \qquad r \geqq 0.$$

Since $a+\gamma(N-2) > N+2$ by assumption (II), ε can be selected small enough and $R \geqq R_0$ large enough that $-\Delta(v-u) \geqq 0$ for all $r \geqq R$. The conclusion then follows as in case (I) by a sufficiently large choice of $A > 0$.

(III) For $b(r) \equiv 0$, the solution $u(r)$ in Theorem 2 satisfies $-\Delta u = \lambda h(r)u^{\gamma}$, where $h(r) = p(r)g(u(r))$. Since $u(r) \to 0$ as $r \to \infty$, conditions (A) and (B) imply that there is a constant $r_1 \geqq r_0$ such that

$$C_1 C_3 r^{-a} \leqq h(r) \leqq C_2 C_4 r^{-a} \quad \text{for all } r \geqq r_1. \tag{24}$$

In view of (1) and the decay estimate in Theorem 2, there exists a constant $C > 0$ such that

$$0 < -[r^{N-1}u'(r)]' = \lambda r^{N-1}p(r)f(u(r)) \leqq Cr^{\alpha}, \qquad r \geqq r_1, \tag{25}$$

where $\alpha = -a-\gamma(N-2)/2+N-1$. Integration over (r_1, r) implies that

$$0 < -u'(r) \leqq K_1 r^{1-N} + K_2 r^{\beta}, \qquad r > r_1 \tag{26}$$

for some constants $K_1>0$, $K_2>0$, $\beta=\alpha-N+2$. Since $\gamma(N-2)>N-2a+2$ for $0\leqq a<2$ by (4), it follows that $\beta<-N/2$, from which $u'(r)=O(r^{-N/2})$ as $r\to\infty$ by (26). Then assumption (C) yields the asymptotic estimate

$$(\log h)'(r)=(\log p)'(r)+(\log g)'(u(r))u'(r)=O(r^{-1}) \tag{27}$$

as $r\to\infty$. In view of (24) and (27), Theorem 3.6(iii) of [6] is applicable for γ in our range (4), $0\leqq a<2$, implying that either $u(r)=O(r^{2-N})$ as $r\to\infty$ or $u(r)\geqq Cr^{(a-2)/(\gamma-1)}$ for large r. However, the second alternative is impossible since $u\in L^{2N/(N-2)}(\mathbf{R}^N)$ by Theorem 2 and

$$\left(\frac{a-2}{\gamma-1}\right)\left(\frac{2N}{N-2}\right)>-N \quad \text{by (4).}$$

If we use the estimate $u(r)=O(r^{2-N})$ instead of the estimate stated in Theorem 2, then α in (25) is replaced by

$$\alpha=-a-\gamma(N-2)+N-1<a-3<-1,$$

implying the boundedness of $r^{N-1}u'(r)$ for $r\geqq r_1$.

REFERENCES

[1] H. BERESTYCKI AND P. L. LIONS, *Une méthode locale pour l'existence de solutions positives de problèmes semi-linéaires elliptiques dans* $\mathbf{R}^N$, J. Analyse Math., 38 (1980), pp. 144-187.

[2] ———, *Nonlinear scalar field equations* I, II, Arch. Rational Mech. Anal., 82 (1983), pp. 313-375.

[3] M. S. BERGER AND M. SCHECHTER, *Embedding theorems and quasi-linear elliptic boundary value problems for unbounded domains*, Trans. Amer. Math. Soc., 172 (1972), pp. 261-278.

[4] Y. FURUSHO, *Positive solutions of linear and quasilinear elliptic equations in unbounded domains*, Hiroshima Math. J., 15 (1985), pp. 173-220.

[5] B. GIDAS, W.-M. NI, AND L. NIRENBERG, *Symmetry of positive solutions of nonlinear elliptic equations in* $\mathbf{R}^n$, Math. Anal. Appl. A, Adv. Math. Suppl. Stud., 7A (1981), pp. 369-402.

[6] B. GIDAS AND J. SPRUCK, *Global and local behaviour of positive solutions of nonlinear elliptic equations*, Comm. Pure Appl. Math., 34 (1981), pp. 525-598.

[7] N. KAWANO, *On bounded entire solutions of semilinear elliptic equations*, Hiroshima Math. J., 14 (1984), pp. 125-158.

[8] T. KUSANO AND M. NAITO, *Positive entire solutions of superlinear elliptic equations*, Hiroshima Math. J., 16 (1986), pp. 361-366.

[9] T. KUSANO AND S. OHARU, *On entire solutions of second order semilinear elliptic equations*, J. Math. Anal. Appl., 113 (1986), pp. 123-135.

[10] E. S. NOUSSAIR AND C. A. SWANSON, *Oscillation theory for semilinear Schrödinger equations and inequalities*, Proc. Roy. Soc. Edinburgh Sect. A, 75 (1975/76), pp. 67-81.

[11] ———, *Global positive solutions of semilinear elliptic problems*, Pacific J. Math., 115 (1984), pp. 177-192.

[12] ———, *Positive solutions of semilinear elliptic problems in unbounded domains*, J. Differential Equations, 57 (1985), pp. 349-372.

[13] W. A. STRAUSS, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys. 55 (1977), pp. 149-162.

[14] C. A. STUART, *Bifurcation for Dirichlet problems without eigenvalues*, Proc. London Math. Soc. (3), 45 (1982), pp. 169-192.

[15] J. S. W. WONG, *On the generalized Emden-Fowler equation*, SIAM Rev., 17 (1975), pp. 339-360.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1344-1365, November 1989

005

HYPERBOLIC-PARABOLIC SINGULAR PERTURBATIONS FOR QUASILINEAR EQUATIONS*

BENJAMIN F. ESHAM, JR.† AND RICHARD J. WEINACHT‡

Abstract. A hyperbolic-parabolic singular perturbation problem is considered for a quasilinear wave equation that arises in one-dimensional nonlinear elasticity. An initial boundary value problem is treated, in which there is an initial layer at $t=0$. It is proved that the solution of the reduced problem approximates the solution of the full problem uniformly on sets bounded in the time direction. An initial-layer corrector and an additional outer expansion term are provided, which yield a uniform $\mathcal{O}(\varepsilon^2)$ approximation.

Key words. singular perturbations, hyperbolic, parabolic, quasilinear, nonlinear wave equations, nonlinear elasticity, nonlinear heat equations

AMS(MOS) subject classifications. 35B25, 35B45, 35B65, 35C20, 35K20, 35L05, 35L20, 73C50, 73D35, 73K03

1. Introduction.

In this paper we consider the singular perturbation problem

$$\varepsilon^2 u_{tt} + u_t - [\sigma(u_x)]_x = f(x,t), \qquad x\in(0,1), \quad t>0, \tag{1.1}$$

$$u(x,0;\varepsilon)=\phi(x),\ \varepsilon u_t(x,0;\varepsilon)=\psi(x), \qquad x\in[0,1], \tag{1.2}$$

$$u(0,t;\varepsilon)=u(1,t;\varepsilon)=0, \qquad t\geqq 0, \tag{1.3}$$

that arises in one-dimensional nonlinear elasticity. For small positive ε, the function u can be interpreted as the nondimensional displacement of the planar oscillations of a (nonlinear) hyperelastic string that is oscillating in a highly viscous medium. The scaling leading to (1.1)-(1.3) in the linear case ($\sigma(z)=z$) is given in [8].

For nonzero ε, (1.1) was considered in an often-referenced but unpublished 1975 work of Nishida (see Nishida [11] where the Cauchy problem is treated). Equation (1.1), as well as corresponding linear and semilinear versions [5], [9], has also been considered as a model for heat conduction to avoid the paradox of infinite speeds of propagation for the parabolic problem (see the bibliographies of [11] and [12]).

For $\varepsilon=0$ we have, in place of (1.1)-(1.3), the well-posed parabolic initial boundary value problem

$$U_t - [\sigma(U_x)]_x = f(x,t), \qquad x\in(0,1), \quad t>0, \tag{1.4}$$

$$U(x,0)=\phi(x), \qquad x\in[0,1], \tag{1.5}$$

$$U(0,t)=U(1,t)=0, \qquad t\geqq 0, \tag{1.6}$$

where the initial condition on u_t in (1.2) has been omitted. Thus there is a nonuniformity near $t=0$, where an initial layer is present.

Our main result (Theorem 5) is the following: on the closure of any rectangular region $Q\doteq(0,1)\times(0,t_0)$ in xt-space on which (1.1)-(1.3) has the solution u, with properties given by Theorem 2, and on which (1.4)-(1.6) has solution U, with properties given by Theorem 3, we have

$$u(x,t;\varepsilon)=U(x,t)+\varepsilon U_1(x,t)+\varepsilon V(x,t/\varepsilon^2)+\mathcal{O}(\varepsilon^2) \quad \text{as } \varepsilon\to 0^+ \tag{1.7}$$

* Received by the editors December 30, 1987; accepted for publication March 30, 1989.

† Department of Mathematical Sciences, Virginia Commonwealth University, 1015 West Main Street, Richmond, Virginia 23284.

‡ Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716.

uniformly on $\bar{Q}$. In (1.7) the function U_1 is a first-order corrector to the outer solution U and the function V is an initial-layer corrector, which we describe in greater detail later in this introduction.

In a previous paper [6], we considered, in place of (1.1), the nonlinear wave equation

$$\varepsilon^2 u_{tt} + u_t - g\left\{\int_0^1 u_x^2(x,t)\,dx\right\} u_{xx} = f(x,t) \tag{1.8}$$

with a nonlocal *scalar nonlinearity*. We refer to [6] for further information about singular perturbations for hyperbolic equations and related works, mentioning here only the recent work of Benaouda and Madaune-Tort [1] on semilinear hyperbolic-parabolic singular perturbations. The required estimates for (1.8) are somewhat less complicated than the present case, and it was the insight gained there that enabled us to handle (1.1).

A key point in our analysis here and in [6] consists of higher-order ε-weighted energy estimates. Energy estimates for hyperbolic singular perturbation problems seem to have been introduced by deJager [3] (see also deJager and Geel [4] and Geel [7]). Higher-order energy estimates were also used by Nishida [11] and have been used effectively in recent years for problems in continuum mechanics (see, for example, Dafermos and Nohel [2] and Slemrod [13]) in more complicated situations than that considered here, but not concerning singular perturbations.

Several points should be emphasized. Quite a bit of smoothness is required for our result. This is due in part to our method, but it should also be observed that the uniform nature of the result (1.7), with correctors present, requires smoothness beyond that needed for the existence of weak solutions of (1.1)-(1.3).

With the extra smoothness, we prove the local existence-uniqueness of $u = u_\varepsilon$ in (1.1)-(1.3) for any ε in some interval $(0, \varepsilon_0]$ and U in (1.4)-(1.6). Nishida [11] obtains global smooth solutions for the Cauchy problem for (1.1), provided a small data assumption is satisfied. Our asymptotic result is of an ideal type in that it is valid uniformly on any rectangular domain for which the above-mentioned u and U are valid.

Our methods do not use any tools that are specifically tied to one space dimension. It is therefore expected that our results can be carried over to more space variables, but with an attendant increase in regularity assumptions.

Both $u = u_\varepsilon$ and U are treated by the Schauder technique. A related linear problem for u is considered in § 2 and the nonlinear hyperbolic problem in § 3. Section 4 contains the corresponding result for U as well as the (linear) problem for the corrector U_1. There also the elementary problems for the initial-layer correctors are solved explicitly. The main result appears in § 5.

Let us now close this introduction by describing the problems that characterize the higher-order correction terms. The natural Ansatz [8], [9]

$$u(x,t;\varepsilon) = U(x,t) + \sum_{n=1}^{N} [U_n(x,t) + V_{n-1}(x, t/\varepsilon^2)]\varepsilon^n, \tag{1.9}$$

where $\tau \doteq t/\varepsilon^2$ is the stretched time variable, leads in the usual way to the following problems. For the outer solution U_1 we have the linear parabolic problem (P_1):

$$U_{1_t} - [\sigma'(U_x) U_{1_x}]_x = 0, \qquad (x,t) \in Q, \tag{1.10}$$

$$U_1(x,0) = \psi(x), \qquad x \in [0,1], \tag{1.11}$$

$$U_1(0,t) = U_1(1,t) = 0, \qquad t \in [0, t_0]. \tag{1.12}$$

For the initial-layer correctors V_n, we have simple linear ODE problems. For the initial-layer corrector $V \equiv V_0$ of lowest order, we have the problem $(\tilde{P}_0)$:

$$\ddot{V} + \dot{V} = 0, \qquad \tau > 0, \quad x \in (0, 1), \tag{1.13}$$

$$\dot{V}(x, 0) = \psi(x), \qquad x \in [0, 1], \tag{1.14}$$

$$V(0, \tau) = V(1, \tau) = 0, \qquad \tau > 0, \tag{1.15}$$

together with the matching condition

$$\lim_{\tau \to \infty} V(x, \tau) = 0, \qquad x \in [0, 1]. \tag{1.16}$$

Here and below, $\partial/\partial\tau$ is denoted by "$\cdot$".

The higher-order initial-layer correctors are determined by very similar problems. In particular, for V_1 we have the problem $(\tilde{P}_1)$:

$$\ddot{V}_1 + \dot{V}_1 = 0, \qquad \tau > 0, \quad x \in (0, 1), \tag{1.17}$$

$$\dot{V}_1(x, 0) = -U_t(x, 0), \qquad x \in [0, 1], \tag{1.18}$$

$$V_1(0, \tau) = V_1(1, \tau) = 0, \qquad \tau > 0, \tag{1.19}$$

with the same matching condition as for V,

$$\lim_{\tau \to \infty} V_1(x, \tau) = 0, \qquad x \in [0, 1]. \tag{1.20}$$

The problems for the outer solutions U_n, $n \geqq 2$, are found to be *nonhomogeneous* linear parabolic problems, but are similar to that for U_1. The same is true for the ODEs for the initial-layer correctors V_n, $n \geqq 2$. The proof of uniform asymptotic validity for the Nth order expansion (1.9) follows from a priori estimates like (5.1). The method should be clear from our treatment here of the case $N = 1$.

2. A related linear problem. As a tool in treating (1.1)-(1.3), we consider the related linear hyperbolic problem $(\mathscr{P}_\varepsilon)$:

$$L_\varepsilon[u] \doteq \varepsilon^2 u_{tt} + u_t - a(x, t; \varepsilon) u_{xx} = f(x, t), \qquad x \in (0, 1), \quad 0 < t < t_0, \tag{2.1}$$

$$u(x, 0; \varepsilon) = \phi(x), \quad \varepsilon u_t(x, 0; \varepsilon) = \psi(x), \qquad x \in [0, 1], \tag{2.2}$$

$$u(0, t; \varepsilon) = u(1, t; \varepsilon) = 0, \qquad t \in [0, t_0], \tag{2.3}$$

under hypotheses on the data $\{\phi, \psi, f\}$ and $a(x, t; \varepsilon)$ appropriate for the subsequent use of the Schauder technique in dealing with the nonlinear problem (1.1)-(1.3) on $Q \doteq (0, 1) \times (0, t_0)$, which we refer to as (P_ε).

Specifically, we make the following hypotheses for arbitrary ε_0 in $(0, 1]$ and $t_0 > 0$. Assume

$$a \in H^3(Q), \tag{a.1}$$

and, for positive constants a_0 and a_1, independent of $(x, t; \varepsilon) \in \bar{Q} \times (0, \varepsilon_0]$

$$0 < a_0 \leqq a(x, t; \varepsilon), \tag{a.2}$$

$$\int_0^{t_0} \|a_t(\cdot, t; \varepsilon)\|_{L^\infty(0,1)}\, dt \leqq a_1, \tag{a.3}$$

$$L^\infty(0, 1) \text{ norms of } \{a(\cdot, 0; \varepsilon), a_x(\cdot, 0; \varepsilon), \varepsilon a_t(\cdot, 0; \varepsilon)\} \leqq a_1, \tag{a.4}$$

$$L^2(0, 1) \text{ norms of } \{a_{xx}(\cdot, 0; \varepsilon), \varepsilon a_{tx}(\cdot, 0; \varepsilon), \varepsilon^3 a_{tt}(\cdot, 0; \varepsilon)\} \leqq a_1, \tag{a.5}$$

$$L^2(Q) \text{ norms of } \{a_x, a_{xx}, a_{xxx}, \varepsilon a_{tx}, \varepsilon a_{txx}, \varepsilon^3 a_{tt}, \varepsilon^3 a_{ttx}, \varepsilon^5 a_{ttt}\} \leqq a_1. \tag{a.6}$$

Moreover, we impose the compatibility condition

$$a_x(0, t; \varepsilon) = a_x(1, t; \varepsilon) = 0, \qquad t \in [0, t_0], \quad \varepsilon \in (0, \varepsilon_0]. \tag{a.7}$$

For the data, we require that

$$\phi \in H^4(0, 1), \qquad \phi = \phi'' = 0 \quad \text{at } x = 0, 1, \tag{ϕ.1}$$

$$\psi \in H^3(0, 1), \qquad \psi = \psi'' = 0 \quad \text{at } x = 0, 1, \tag{ψ.1}$$

$$f \in H^3(Q), \qquad f = f_{xx} = 0 \quad \text{at } x = 0, 1. \tag{f.1}$$

In what follows, we will often abbreviate $L^\infty(0, t_0; L^2(0, 1))$ as $L^\infty(L^2)$.

THEOREM 1. *If a, ϕ, ψ, and f satisfy the above hypotheses, the problem $(\mathcal{P}_\varepsilon)$ has a unique classical solution $u \in C^2(\bar{Q})$ with $D^\alpha u \in L^\infty(L^2)$, $|\alpha| \leqq 4$, such that each of the following norms is bounded by a constant independent of ε in $(0, \varepsilon_0]$:*

(r.1) *$L^\infty(Q)$ norms of* $\{u, u_x, u_{xx}, u_{xxx}, \varepsilon u_t, \varepsilon u_{tx}, \varepsilon u_{txx}, \varepsilon^3 u_{tt}, \varepsilon^3 u_{ttx}, \varepsilon^5 u_{ttt}\}$,

(r.2) *$L^\infty(L^2)$ norms of* $\{u_{xxxx}, \varepsilon u_{txxx}, \varepsilon^3 u_{ttxx}, \varepsilon^5 u_{tttx}, \varepsilon^7 u_{tttt}\}$,

(r.3) *$L^2(Q)$ norms of* $\{u_t, u_{tx}, u_{txx}, u_{txxx}, \varepsilon^2 u_{tt}, \varepsilon^2 u_{ttx}, \varepsilon^2 u_{ttxx}, \varepsilon^4 u_{ttt}, \varepsilon^4 u_{tttx}, \varepsilon^6 u_{tttt}\}$.

Proof. Imposing on the m-term Faedo-Galerkin approximation

$$u_m \equiv u_m(x, t; \varepsilon) \doteq \sum_{k=1}^{m} A_{km}(t; \varepsilon) \sin k\pi x, \tag{2.4}$$

the orthogonality conditions with $L^2(0, 1)$ inner product $\langle \cdot, \cdot \rangle$

$$\langle L_\varepsilon u_m - f_m, \sin l\pi x \rangle = 0, \qquad l = 1, 2, \cdots, m, \tag{2.5}$$

as well as projected initial conditions, we are led to the following initial-value problem for the coefficients $\{A_{lm}\}_{l=1}^m$

$$\varepsilon^2 A_{lm}'' + A_{lm}' + 2 \sum_{k=1}^{m} k^2 \pi^2 a_{lk}(t; \varepsilon) A_{km} = f_l(t), \tag{2.6}$$

$$A_{lm}(0; \varepsilon) = 2\langle \phi, \sin l\pi x \rangle, \qquad \varepsilon A_{lm}'(0; \varepsilon) = 2\langle \psi, \sin l\pi x \rangle, \tag{2.7}$$

for $l = 1, 2, \cdots, m$. The functions $a_{lk}(t; \varepsilon)$, $f_m(x, t)$ and $f_l(t)$ are defined by

$$a_{lk}(t; \varepsilon) = \langle a(x, t; \varepsilon) \sin k\pi x, \sin l\pi x \rangle$$

and

$$f_m(x, t) = \sum_{l=1}^{m} f_l(t) \sin l\pi x = 2 \sum_{l=1}^{m} \langle f(x, t), \sin l\pi x \rangle \sin l\pi x. \tag{2.8}$$

From elementary regularity results for systems of linear ODEs, it follows that the unique solution $\{A_{lm}\}_{l=1}^m$ of (2.6)-(2.7) belongs to $H^5(0, t_0)$, so that u_m, $m = 1, 2, \cdots$, belongs to $C^4(\bar{Q}) \cap H^5(Q)$ and, in fact, is analytic as a function of x for fixed t.

To show the convergence of the u_m to the solution u of $(\mathcal{P}_\varepsilon)$, we use energy estimates of higher order with appropriate ε-weights. Let

$$E(t) \doteq \frac{1}{2} \int_0^1 \left\{ \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2 + 2} |D^\alpha u_{m_t}|^2 + a(x, t; \varepsilon) \sum_{\alpha_1 = 0}^{3} |D_x^{\alpha_1} u_{m_x}|^2 \right\} dx \tag{2.9}$$

where the usual multi-index $\alpha=(\alpha_1,\alpha_2)$ is used: $D^\alpha=D_x^{\alpha_1}D_t^{\alpha_2}$. The term in $E(t)$ corresponding to $\alpha=(0,0)$ is the usual energy at time t of the vibrating string governed by (2.1).

Our basic energy estimate is

$$\begin{aligned}E(t)+\frac{1}{4}\int_0^t\int_0^1\sum_{0\leqq|\alpha|\leqq3}\varepsilon^{4\alpha_2}|D^\alpha u_{m_t}|^2\,dx\,ds\\ \leqq\left(F(0)+\frac{1}{2}\|f_m\|^2_{H^3(Q)}\right)\exp\left\{\int_0^t K(s)\,ds\right\}\end{aligned}\tag{2.10}$$

where

$$F(t)=\frac{1}{2}\sum_{0\leqq|\alpha|\leqq3}\varepsilon^{4\alpha_2}\int_0^1\{\varepsilon^2|D^\alpha u_{m_t}|^2+a(x,t;\varepsilon)|D^\alpha u_{m_x}|^2\}\,dx,\tag{2.11}$$

$$K(s)=K_0\left(1+\sup_{0\leqq x\leqq1}|a_t(x,s)|+\int_0^1\left\{\sum_{1\leqq|\alpha|\leqq2}\varepsilon^{4\alpha_2+2}|D^\alpha a_t|^2+\sum_{\alpha_1=0}^{2}|D_x^{\alpha_1}a_x|^2\right\}dx\right).\tag{2.12}$$

The constant K_0 is independent of m and ε.

The estimate (2.10) is established in the appendix, where it is also shown how appropriate estimates of f_m in the $H^3(Q)$ norm and of $F(0;\varepsilon)$ lead to the inequality

$$\begin{aligned}\sup_{0\leqq t\leqq t_0}\int_0^1\left(\sum_{0\leqq|\alpha|\leqq3}\varepsilon^{4\alpha_2+2}|D^\alpha u_{m_t}|^2+\sum_{\alpha_1=0}^{3}|D^\alpha u_{m_x}|^2\right)dx\\ +\sum_{0\leqq|\alpha|\leqq3}\varepsilon^{4\alpha_2}\int_0^{t_0}\int_0^1|D^\alpha u_{m_t}|^2\,dx\,ds\leqq C\end{aligned}\tag{2.13}$$

with constant C independent of m and ε.

From this energy inequality, we see that each of the sequences $\{D^\alpha u_m\}_{m=1}^\infty$, for $0\leqq|\alpha|\leqq4$, is bounded in $L^\infty(L^2)$ and so contains a weak* convergent subsequence. If u denotes the weak* limit of the subsequence for $\alpha=(0,0)$, then for $0<|\alpha|\leqq4$, there is a subsequence $\{D^\alpha u_{m_j}\}_{j=1}^\infty$ that has limit $D^\alpha u$, the αth weak $L^2(Q)$ derivative of u. Hence $D^\alpha u\in L^\infty(L^2)$, $0\leqq|\alpha|\leqq4$, and $u\in H^4(Q)$, so that $u\in C^2(\bar{Q})$ by the Sobolev embedding theorem.

To see that the limit u satisfies (2.1), we merely let m_j tend to infinity in

$$\int_0^{t_0}\int_0^1(Lu_{m_j}-f_{m_j})\mu(t)\sin l\pi x\,dx\,dt=0$$

for arbitrary $\mu\in L^2(0,t_0)$. It follows by a standard density argument that (2.1) is satisfied almost everywhere in Q and hence everywhere in $\bar{Q}$, since all the functions involved are continuous in $\bar{Q}$.

To see that the weak* limit u satisfies the initial condition $u(x,0)=\phi(x)$, we proceed in a similar and familiar way. We note that for $\alpha\in L^2(0,1)$ and $\beta\in C^1[0,t_0]$, such that $\beta(0)=1$ and $\beta(t_0)=0$,

$$\int_0^{t_0}\int_0^1 u_{m_j}(x,t)\alpha(x)\beta'(t)\,dx\,dt\to\int_0^{t_0}\int_0^1 u(x,t)\alpha(x)\beta'(t)\,dx\,dt,$$

and

$$\int_0^{t_0}\int_0^1 u_{m_j,t}(x,t)\alpha(x)\beta(t)\,dx\,dt\to\int_0^{t_0}\int_0^1 u_t(x,t)\alpha(x)\beta(t)\,dx\,dt,$$

from which

$$\int_0^1 u_{m_j}(x,0)\alpha(x)\,dx \to \int_0^1 u(x,0)\alpha(x)\,dx,$$

and the condition follows since $u_m(x,0)=\phi_m(x)\to\phi(x)$ in $L^2(0,1)$. A similar argument shows that $\varepsilon u_t(x,0)=\psi(x)$ and that $u=u_{xx}=0$ when $x=0,1$.

It is conceivable that there are distinct subsequences of $\{u_m\}$ that converge to different limits, and thus generate different solutions of $(\mathscr{P}_\varepsilon)$. However, global uniqueness for solutions of $(\mathscr{P}_\varepsilon)$ follows from an easy energy integral argument. In particular, if u and $\bar{u}$ are classical solutions of $(\mathscr{P}_\varepsilon)$, the function $z \doteq u-\bar{u}$ satisfies the corresponding completely homogeneous IBVP. With

$$\mathbb{E}(t)\doteq\frac{1}{2}\int_0^1 [\varepsilon^2 z_t^2 + a(x,t;\varepsilon)z_x^2]\,dx,$$

we quickly derive the identity and bound

$$\begin{aligned}\mathbb{E}(t)+\int_0^t\int_0^1 z_t^2\,dx\,ds &= \int_0^t\int_0^1\left[\frac{1}{2}a_t z_x^2 - a_x z_x z_t\right]dx\,ds\\ &\leqq \frac{1}{2}\int_0^t\int_0^1 z_t^2\,dx\,ds + K_1\int_0^1 \mathbb{E}(s)\,ds,\end{aligned}$$

where

$$K_1 \doteq a_0^{-1}\left\{\sup_{\bar{Q}}|a_t| + \sup_{\bar{Q}}|a_x|^2\right\}.$$

Gronwall's lemma immediately implies

$$\mathbb{E}(t)\equiv 0 \Rightarrow z\equiv 0 \Rightarrow u\equiv\bar{u}.$$

It follows that each subsequence of $\{u_m\}$ must converge to the unique solution u of $(\mathscr{P}_\varepsilon)$.

The assertions (r.2) and (r.3) of Theorem 1 on the order ε behavior of the various derivatives of u follow directly from the inequality (2.13) when we use the lower semicontinuity property of weak* limits. With these established, (r.1) follows by means of the Sobolev embedding theorem. This completes the proof of Theorem 1. □

3. The quasilinear hyperbolic equation. With Theorem 1 of § 2 established, the way has been prepared to approach the questions of existence and uniqueness for the quasilinear problem (P_ε) by means of the Schauder technique and the Banach contraction mapping theorem. Indeed, the fixed point will also be seen to inherit the regularity and order ε behavior built into the linear problem $(\mathscr{P}_\varepsilon)$. We proceed then with the following theorem.

THEOREM 2. *Let f, ϕ and ψ satisfy hypotheses* (f.1), (ϕ.1), *and* (ψ.1), *respectively. Let $\sigma\in C^4(\mathbb{R})$ with $0<\sigma_0\leqq\sigma'(z)$ for all $z\in\mathbb{R}$. Then the quasilinear hyperbolic problem (P_ε) consisting of* (1.1)-(1.3) *on Q has a unique classical solution $u\in C^2(\bar{Q})$ with $D^\alpha u\in L^\infty(L^2)$, $|\alpha|\leqq 4$, such that conclusions* (r.1), (r.2), *and* (r.3) *of Theorem* 1 *hold. The interval $[0,t_0]$ may be small, but it is independent of ε in $(0,\varepsilon_0]$.*

Proof. Let $\mathscr{Y}=\mathscr{Y}(\varepsilon,t_0)$ be the complete metric space

$$\mathscr{Y}=\{v\in L^\infty(0,t_0;H_0^1(0,1))\,|\,v_t\in L^\infty(0,t_0;L^2(0,1))\}$$

with ε-dependent metric $d=d_\varepsilon$ defined by

$$d^2(v,\bar{v})\doteq\sup_{0\leqq t\leqq t_0}\int_0^1[(v_x-\bar{v}_x)^2+\varepsilon^2(v_t-\bar{v}_t)^2]\,dx.$$

For suitable functions $v = v(x, t; \varepsilon)$, with $\varepsilon \in (0, \varepsilon_0]$, let

$$\delta(v) \doteq \max_{0 \leqq |\alpha| \leqq 3} \left\{ \sup_{0 \leqq t \leqq t_0} \int_0^1 \varepsilon^{4\alpha_2+2} |D^\alpha v_t|^2 \, dx, \right.$$

$$\left. \sup_{0 \leqq t \leqq t_0} \int_0^1 |D_x^{\alpha_1} v_x|^2 \, dx, \varepsilon^{4\alpha_2} \int_0^{t_0} \int_0^1 |D^\alpha v_t|^2 \, dx \, ds \right\}.$$

Let $\mathscr{X} \equiv \mathscr{X}(M, t_0, \varepsilon)$ be the subset of $\mathscr{Y}$ consisting of functions $v = v(x, t; \varepsilon)$ such that for given positive M

$$\{v \in H^4(Q): D^\alpha v \in L^\infty(0, t_0; L^2(0, 1)), |\alpha| < 4, \text{ and } \delta(v) \leqq M^2\}$$

and

$$v(0, t; \varepsilon) = v(1, t; \varepsilon) = v_{xx}(0, t; \varepsilon) = v_{xx}(1, t; \varepsilon) = 0,$$

$$v(x, 0; \varepsilon) = \phi(x), \qquad \varepsilon v_t(x, 0; \varepsilon) = \psi(x),$$

$$\varepsilon^2 v_{tt}(x, 0; \varepsilon) = f(x, 0) + \sigma'(\phi'(x))\phi''(x) - \psi(x)/\varepsilon.$$

The set $\mathscr{X}$ is a closed subset of $\mathscr{Y}$. Indeed, let $\{v_n\}$ be a sequence in $\mathscr{X}$ that converges to $\tilde{v} \in \mathscr{Y}$ in the metric on $\mathscr{Y}$. Since $L^\infty(L^2)$ is the dual space of $L^1(L^2)$, any bounded set in $L^\infty(L^2)$ is sequentially compact in the weak* topology of $L^\infty(L^2)$. Since $\delta(v_n) \leqq M^2$, $n = 1, 2, \cdots$, each of the sequences $\{D^\alpha v_n\}$, $|\alpha| \leqq 4$ is bounded in $L^\infty(L^2)$ and so has a subsequence that converges to an element of $L^\infty(L^2)$, $D^\alpha v_{n_j} \to w^\alpha$, weak*. If v is the limit function corresponding to $\alpha = (0, 0)$, it follows easily from definitions that $w^\alpha = D^\alpha v$, the αth weak L^2 derivative of v. Also, by the lower semicontinuity property of weak* sequential convergence,

$$\|\varepsilon^{2\alpha_2+1} D^\alpha v_t\|_{L^\infty(L^2)} \leqq \varliminf_{n \to \infty} \|\varepsilon^{2\alpha_2+1} D^\alpha v_{n_t}\|_{L^\infty(L^2)} \leqq M,$$

$$\|D_x^{\alpha_1} v_x\|_{L^\infty(L^2)} \leqq \varliminf_{n \to \infty} \|D_x^{\alpha_1} v_{n_x}\|_{L^\infty(L^2)} \leqq M,$$

for $|\alpha| \leqq 3$, and by a similar argument involving weak convergence in $L^2(Q)$,

$$\|\varepsilon^{2\alpha_2} D^\alpha v_t\|_{L^2(Q)} \leqq M, \qquad |\alpha| \leqq 3.$$

Hence we conclude that $\delta(v) \leqq M^2$. The conditions on ∂Q are seen to hold for v by arguments similar to those in Theorem 1. Since $v_n \to \tilde{v}$ with respect to the metric d, we clearly have $v_n \to \tilde{v} \in \mathscr{Y}$ weakly in $L^2(Q)$. On the other hand, a subsequence (v_{n_j}) has been shown to converge to $v \in \mathscr{X}$ weakly in $L^2(Q)$. By uniqueness of weak limits, each subsequence must converge to $v = \tilde{v}$. Hence $\tilde{v} \in \mathscr{X}$ and $\mathscr{X}$ is closed.

For a given v in $\mathscr{X}$, consider the linear IBVP (2.1)–(2.3), with

$$a(x, t; \varepsilon) \doteq \sigma'(v_x(x, t; \varepsilon)).$$

It is a straightforward matter to verify that this function satisfies hypotheses (a.1)–(a.7) of Theorem 1. Hence, for each v in $\mathscr{X}$, we obtain a unique solution u of $(\mathscr{P}_\varepsilon)$ that enjoys all the properties guaranteed by Theorem 1. Denote this solution map by $\mathbb{S}$: $u = \mathbb{S}v$.

Seeking fixed points of $\mathbb{S}$, we first show that $\mathbb{S}$ maps $\mathscr{X}$ into itself for appropriately chosen M and t_0. For v in $\mathscr{X}$ the energy inequality (2.10) implies, by the lower semicontinuity property of weak* limits,

$$\delta(u) \leqq \left(k_0[\|\phi\|_{H^4(0,1)}^2 + \|\psi\|_{H^3(0,1)}^2 + \|f(x, 0)\|_*^2] + \frac{1}{2}\|f\|_{H^3(Q)}^2 \right) \exp\left\{ \int_0^{t_0} K(s) \, ds \right\},$$

where $K(s)$ is given by (2.12) and

$$\|f(x,0)\|_*^2 \doteq \sum_{|\alpha|\leqq 2} \int_0^1 |D^\alpha f(x,0)|^2\,dx.$$

We now choose $M>1$ so large that the data at $t=0$ is bounded above as

$$k_0[\|\phi\|^2_{H^4(0,1)}+\|\psi\|^2_{H^3(0,1)}+\|f(x,0)\|_*^2]\leqq \tfrac{1}{2}M,$$

and then choose t_0 sufficiently small so that $\|f\|^2_{H^3(Q)}<\frac{1}{2}M$. In considering the exponential factor, it is important to note that

$$\begin{aligned}\int_0^{t_0} \sup_{0\leqq x\leqq 1} |a_t|\,dt &= \int_0^{t_0} \sup_{0\leqq x\leqq 1} |(t_0^{-1/4}\sigma''(v_x))(t_0^{1/4}v_{tx})|\,dt\\ &\leqq \frac{1}{2}\sup_{\bar Q}|\sigma''(v_x)|^2\sqrt{t_0}+\frac{1}{2}\sqrt{t_0}\int_0^{t_0}\int_0^1 (v_{tx}^2+v_{txx}^2)\,dx\,dt\\ &\leqq k(M)\sqrt{t_0},\end{aligned}$$

where k depends on M, but not on ε. It follows from the assumed order ε behavior of the derivatives of v in $\mathscr{X}$ that

$$\int_0^1 \left\{\sum_{1\leqq|\alpha|\leqq 2}\varepsilon^{4\alpha_2+2}|D^\alpha a_t|^2+\sum_{\alpha_1=0}^{2}|D_x^{\alpha_1}a_x|^2\right\}dx\leqq k(M)$$

holds for a constant depending on M but independent of ε. Hence

$$\exp\left\{\int_0^{t_0}K(s)\,ds\right\}\leqq \exp\{k(M)\sqrt{t_0}\}\leqq M,$$

provided t_0 is chosen sufficiently small. This implies that $\delta(u)\leqq M^2$. The remaining subsidiary conditions in $\mathscr{X}$ follow from the fact that u is a classical solution of $(\mathscr{P}_\varepsilon)$ on $\bar Q$. Hence $u\in\mathscr{X}$ and $\mathbb{S}:\mathscr{X}\to\mathscr{X}$.

To show that $\mathbb{S}$ is a contraction on $\mathscr{X}$ with respect to the metric d, let $v,\bar v\in\mathscr{X}$ and $u\doteq\mathbb{S}v$, $\bar u\doteq\mathbb{S}\bar v$. Then $w\doteq u-\bar u$ is a classical solution of the linear initial boundary value problem consisting of the PDE

$$\varepsilon^2 w_{tt}+w_t-\sigma'(v_x)w_{xx}=[\sigma'(v_x)-\sigma'(\bar v_x)]\bar u_{xx}, \tag{3.1}$$

and the conditions (2.2)-(2.3) with $\phi(x)\equiv\psi(x)\equiv 0$. The energy of w defined by

$$\tilde E(t)\doteq\frac{1}{2}\int_0^1[\varepsilon^2 w_t^2+\sigma'(v_x)w_x^2]\,dx \tag{3.2}$$

satisfies

$$\begin{aligned}&\tilde E(t)+\int_0^t\int_0^1 w_t^2\,dx\,ds\\ &\leqq\int_0^t\int_0^1\left\{w_t[(\sigma'(v_x)-\sigma'(\bar v_x))\bar u_{xx}-\sigma''(v_x)v_{xx}w_x]+\frac{1}{2}\sigma''(v_x)v_{tx}w_x^2\right\}dx\,ds,\end{aligned} \tag{3.3}$$

and so by estimates of the right-hand side similar to those in the Appendix, we obtain via Gronwall's lemma

$$\tilde E(t)\leqq C_1 t_0 d^2(v,\bar v)\exp\left\{\int_0^{t_0}\tilde K(s)\,ds\right\},$$

with

$$C_1 \doteq \sup_{\substack{\bar{Q} \\ 0<\theta<1}} |\sigma''(\theta v_x + (1-\theta)\bar{v}_x)\tilde{u}_{xx}|^2,$$

independent of ε, and kernel given by

$$\tilde{K}(s) \doteq \max\{1, \sigma_0^{-1}\}\left(\frac{1}{2}\sup_{\bar{Q}} |\sigma''(v_x)v_{xx}| + \sup_{\bar{Q}} |\sigma''(v_x)|\left\{1 + \int_0^1 (v_{tx}^2 + v_{txx}^2)\, dx\right\}\right).$$

The assumptions on v and our results on u guarantee that the $L^1(0, t_0)$ norm of $\tilde{K}(t)$ is bounded independently of ε in $(0, \varepsilon_0]$. Now

$$d^2(u, \bar{u}) \leqq \max\{1, \sigma_0^{-1}\} \sup_{0 \leqq t \leqq t_0} \tilde{E}(t) \leqq Ct_0 d^2(v, \bar{v}),$$

so by choice of t_0 sufficiently small, the map $\mathbb{S}$ is a contraction as asserted. The Banach contraction mapping theorem implies that $\mathbb{S}$ has a unique fixed point in $\mathscr{X}$, which is seen to be a classical solution of (P_ε). Moreover, there is at most one classical solution $u \in C^2(\bar{Q})$ for this t_0 as shown by the following energy integral argument. Let u and $\tilde{u}$ be classical $C^2(\bar{Q})$ solutions of (P_ε); then (3.1)-(3.3) hold with $v = u$ and $\bar{v} = \tilde{u}$. With

$$\sup_{\bar{Q}} \{|\sigma''(u_x)u_{xx}|, |\sigma''(u_x)u_{tx}|, |\sigma''(\theta u_x + (1-\theta)\tilde{u}_x)\tilde{u}_{xx}|\} \leqq k,$$

we quickly obtain the estimate

$$\tilde{E}(t) \leqq \frac{2k^2 + k}{\sigma_0} \int_0^t \tilde{E}(s)\, ds,$$

and, by Gronwall's lemma, $\tilde{E}(t) \leqq 0$, so that $u = \tilde{u}$ follows immediately. □

4. The reduced problem and higher-order correctors. In this section we consider the reduced problem (P_0) and the problems (P_1), $(\tilde{P}_0)$ and $(\tilde{P}_1)$ introduced in § 1. The initial-layer corrector problems $(\tilde{P}_0)$ and $(\tilde{P}_1)$ readily yield explicit solutions given by

$$V(x, \tau) = -\psi(x)\, e^{-\tau}, \tag{4.1}$$

and

$$V_1(x, \tau) = \{\sigma'(\phi'(x))\phi''(x) + f(x, 0)\}\, e^{-\tau}. \tag{4.2}$$

In § 5 our proof of uniform validity for the expansion will require that V_{xx} and $V_{1_{xx}}$ belong to $L^\infty(Q)$. From (4.1) and (4.2) this will certainly be the case if

$$\sigma \in C^3, \quad \phi^{iv} \in L^\infty(0, 1), \quad \psi'' \in L^\infty(0, 1), \quad f_{xx}(x, 0) \in L^\infty(0, 1).$$

As indicated in (1.4)-(1.6), the reduced problem (P_0) corresponding to $\varepsilon = 0$ consists of the quasilinear parabolic PDE

$$U_t - \sigma'(U_x)U_{xx} = f(x, t), \qquad (x, t) \in Q, \tag{4.3}$$

together with the single initial condition

$$U(x, 0) = \phi(x), \qquad x \in [0, 1], \tag{4.4}$$

and the homogeneous boundary conditions

$$U(0, t) = U(1, t) = 0, \qquad t \in [0, t_0]. \tag{4.5}$$

This initial boundary value problem, of course, has been studied in depth both with regard to weak and classical solutions. Nonetheless the kind of results that we seek

for the regularity of the solution do not seem to be explicitly available. Consequently, we are forced to examine this problem with hypotheses compatible with the related hyperbolic problem (P_ε), and, more importantly, sufficient to guarantee existence of a classical solution of (P_1). We state our result concerning the reduced problem as the following theorem.

THEOREM 3. *Let* $\sigma \in C^7(\mathbb{R})$ *satisfy* $0 < \sigma_0 \leqq \sigma'(z)$ *for all* $z \in \mathbb{R}$.

Assume

$$(\phi.2) \qquad \phi \in H^7(0,1), \quad \phi = \phi'' = \phi^{iv} = \phi^{vi} = 0 \quad \text{at } x = 0, 1,$$

$$(f.2) \qquad f \in H^3(Q), \quad D_x^{2\alpha_1 - 1} f \in H^{4-\alpha_1}(Q), \quad \alpha_1 = 1, 2, 3,$$
$$f = f_{xx} = f_{xxxx} = 0 \quad \text{at } x = 0, 1.$$

Then the quasilinear parabolic problem (P_0) has a unique classical solution $U(x, t) \in C^{2,1}(\bar{Q})$ such that $U_{xxxx} = 0$ at $x = 0, 1$ and

$$(r.4) \qquad U_{tx}, U_{xxx}, U_{txx}, U_{xxxx} \in C(\bar{Q}),$$

$$(r.5) \qquad U_{tt}, U_{ttt}, U_{ttx}, U_{txxx}, U_{ttxx}, U_{xxxxx}, U_{txxxx}, U_{tttx}, U_{ttxxx}, U_{txxxxx}, U_{xxxxxx} \in L^\infty(L^2),$$

$$(r.6) \qquad U_{tttxx}, U_{ttxxxx}, U_{txxxxxx}, U_{xxxxxxx} \in L^2(Q).$$

Proof. In spite of the similarity of the pattern of proof here with that of Theorems 1 and 2, we will give the essential details for the convenience of the interested reader.

(i) *A related linear parabolic problem.* First consider a related linear parabolic problem for which the PDE (4.3) is replaced by

$$(4.6) \qquad \tilde{\mathbb{L}} U \doteq U_t - b(x, t) U_{xx} = f(x, t),$$

but the same initial condition (4.4) and boundary conditions (4.5) are retained. We assume for $b(x, t)$ that

$$(b.1) \qquad 0 < b_0 \leqq b(x, t),$$

$$(b.2) \qquad b_x = b_{xxx} = 0, \qquad \text{at } x = 0, 1,$$

$$(b.3) \qquad b \in H^3(Q), \qquad D_x^{2\alpha_1 - 1} b \in H^{4-\alpha_1}(Q), \quad \alpha_1 = 1, 2, 3.$$

The Faedo-Galerkin approximations

$$U_m(x, t) = \sum_{k=1}^{m} B_{km}(t) \sin k\pi x$$

for the linear problem (4.6) satisfy the orthogonality conditions

$$(4.7) \qquad \langle D^\alpha(\tilde{\mathbb{L}} U_m - f_m), D^\alpha U_m \rangle = 0$$

for all 19 multi-indices in

$$\mathbb{I} = \{\alpha = (\alpha_1, \alpha_2) | 0 \leqq \alpha_1 + 2 \max\{\tfrac{1}{2}, \alpha_2\} \leqq 7\}.$$

We note that assumption (b.2) is needed in deriving (4.7) when $\alpha_1 = 3$ and 5. From (4.7) we derive an integral relation for each $\alpha \in \mathbb{I}$:

$$\frac{1}{2}\int_0^1 |D^\alpha U_m|^2 \, dx + \int_0^t \int_0^1 b(x, s) |D^\alpha U_{m_x}|^2 \, dx \, ds$$
$$= \frac{1}{2}\int_0^1 |D^\alpha U_m(x, 0)|^2 \, dx + \int_0^t \int_0^1 D^\alpha f_m D^\alpha U_m \, dx \, ds$$
$$+ \int_0^t \int_0^1 D^\alpha U_m \left\{ \sum_{\substack{0 \leqq \beta \leqq \alpha \\ 0 < |\beta|}} \binom{\alpha}{\beta} D^\beta b D^{\alpha - \beta} U_{m_{xx}} - b_x D^\alpha U_{m_x} \right\} dx \, ds.$$

To put these results together, we define

$$\mathscr{E}(t) \doteq \frac{1}{2} \sum_{\alpha \in \mathbb{I}} \int_0^1 |D^\alpha U_m|^2 \, dx.$$

Estimates similar to the hyperbolic case appearing in the Appendix result in the inequality

$$\mathscr{E}(t) + \frac{1}{2} b_0 \sum_{\alpha \in \mathbb{I}} \int_0^t \int_0^1 |D^\alpha U_m|^2 \, dx \, ds \leqq \mathscr{E}(0) + \frac{1}{2} \|f_m\|_\star^2 + \int_0^t K_0(s) \mathscr{E}(s) \, ds,$$

where

$$\|f_m\|_\star^2 \doteq \sum_{\alpha \in \mathbb{I}} \int_0^{t_0} \int_0^1 |D^\alpha f_m|^2 \, dx \, ds$$

and

$$K_0(s) = C \left\{ 1 + t_0^{-1/2} + \sum_{\alpha \in \mathbb{I}_0} \|D^\alpha b\|_{L^\infty(Q)} + \int_0^1 \left(\sum_{\alpha \in \mathbb{I}_1} |D^\alpha b|^2 + t_0^{1/2} \sum_{\alpha \in \mathbb{I}_2} |D^\alpha b|^2 \right) dx \right\},$$

with

$$\mathbb{I}_0 = \{\alpha = (\alpha_1, \alpha_2) | 1 \leqq \alpha_1 + 2\alpha_2 \leqq 3\},$$

$$\mathbb{I}_1 = \{\alpha = (\alpha_1, \alpha_2) | 4 \leqq \alpha_1 + 2\alpha_2 \leqq 6, \alpha \neq (6, 0)\},$$

$$\mathbb{I}_2 = [\alpha = (\alpha_1, \alpha_2) | \alpha_1 + 2 \max \{\tfrac{1}{2}, \alpha_2\} = 7];$$

note that $|\mathbb{I}_0| = 5$, $|\mathbb{I}_1| = 9$, and $|\mathbb{I}_2| = 4$. By Gronwall's lemma,

$$(4.8) \quad \mathscr{E}(t) + \frac{1}{2} b_0 \int_0^t \int_0^1 \sum_{\alpha \in \mathbb{I}} |D^\alpha U_{m_x}|^2 \, dx \, ds \leqq \left\{ \mathscr{E}(0) + \frac{1}{2} \|f_m\|_\star^2 \right\} \exp \int_0^{t_0} K_0(s) \, ds.$$

Hypotheses (ϕ.2) and (f.2) are sufficient to guarantee that the right-hand side is bounded by a constant independent of m. To see this, write

$$\mathscr{E}(0) = \frac{1}{2} \sum_{\alpha_1 = 0}^{6} \int_0^1 |D_x^{\alpha_1} U_m(x, 0)|^2 \, dx + \frac{1}{2} \sum_{\substack{\alpha \in \mathbb{I} \\ \alpha_2 > 0}} \int_0^1 |D^\alpha U_m(x, 0)|^2 \, dx.$$

In the first sum, the terms satisfy

$$\int_0^1 |D_x^{\alpha_1} U_m(x, 0)|^2 \, dx = \int_0^1 |\phi_m^{(\alpha_1)}|^2 \, dx \leqq \int_0^1 |\phi^{(\alpha_1)}|^2 \, dx.$$

For the second sum, we note that

$$\langle D^\alpha (\tilde{\mathbb{L}} U_m - f_m), D^\alpha U_{m_t} \rangle = 0, \qquad \alpha \in \mathbb{I}_3,$$

where $\mathbb{I}_3 = \{\alpha = (\alpha_1, \alpha_2) | 0 \leqq \alpha_1 + 2\alpha_2 \leqq 5\}$. Then

$$\int_0^1 |D^\alpha U_m(x, 0)|^2 \, dx \leqq C \int_0^1 \left\{ |D^\alpha f_m(x, 0)|^2 + \sum_{0 \leqq \beta \leqq \gamma} |D^\beta b(x, 0) D^{\gamma - \beta} U_{m_{xx}}(x, 0)|^2 \right\} dx,$$

where $\gamma = (\alpha_1, \alpha_2 - 1)$ for $\alpha \in \mathbb{I}$, $\alpha_2 > 0$. The norm $\|f_m\|_\star$ and $\int_0^1 |D^\alpha f_m(x, 0)|^2 \, dx$, for $\alpha \in \mathbb{I}$, $\alpha_2 > 0$, are bounded independently of m as is done in (A.11) of the Appendix. Hence each of the sequences

$$\{D^\alpha U_m\}_{\alpha \in \mathbb{I}} \text{ is bounded in } L^\infty(L^2),$$

$$\{D^\alpha U_m\}_{\alpha \in \mathbb{I}_4} \text{ is bounded in } L^2(Q).$$

Here $\mathbb{I}_4 = \{\alpha = (\alpha_1, \alpha_2) | \alpha_1 + 2 \max\{\frac{1}{2}, \alpha_2\} = 8\}$. Uniqueness follows from the identity and bound for the difference, $Z \doteq U - \tilde{U}$, of two classical $C^{2,1}(\bar{Q})$ solutions

$$\frac{1}{2}\int_0^1 Z^2\,dx + \int_0^t \int_0^1 bZ_x^2\,dx\,ds = -\int_0^t \int_0^1 b_x Z_x Z\,dx\,ds$$

$$\leqq \frac{b_0}{2}\int_0^t \int_0^1 Z_x^2\,dx\,ds + b_0^{-1} \sup_{\bar{Q}} |b_x|^2 \int_0^t \left\{\frac{1}{2}\int_0^1 Z_x^2\,dx\,ds\right\}$$

by Gronwall's lemma. The remaining portion of the proof for the linear problem is entirely analogous to the final part of the proof of Theorem 1.

(ii) *Contraction mapping and Schauder technique.* A proper setting for Banach's contraction mapping theorem is obtained by defining a complete metric space

$$\mathbb{Z} \doteq L^\infty(0, t_0; H^1(0, 1)),$$

with

$$\|V\|_z^2 \equiv \|V\|_{L^\infty(H^1)}^2 \doteq \sup_{0 \leqq t \leqq t_0} \int_0^1 [V^2 + V_x^2]\,dx,$$

and a closed subset $\mathbb{X}(M, t_0)$ of $\mathbb{Z}$, consisting of functions V on Q with generalized derivatives

$$D^\alpha V \in L^\infty(0, t_0; L^2(0, 1)), \qquad D^\alpha V \in L^2(Q), \quad \alpha \in \mathbb{I},$$

satisfying

$$V(x, 0) = \phi(x), \qquad V = V_{xx} = V_{xxxx} = 0 \quad \text{at } x = 0, 1,$$

and

$$\rho(V) \doteq \max_{\alpha \in \mathbb{I}} \{\|D^\alpha V\|_{L^\infty(L^2)}, \|D^\alpha V_x\|_{L^2(Q)}\} \leqq M.$$

For $V \in \mathbb{X}$, let $\mathbb{T}$ be the solution operator associated with the linear initial boundary value problem

$$U_t - \sigma'(V_x) U_{xx} = f(x, t),$$
$$U(0, t) = U(1, t) = 0,$$
$$U(x, 0) = \phi(x),$$

so that $U = \mathbb{T} V$. The proof is complete if $\mathbb{T}$ is a contraction map on $\mathbb{X}$. With $b(x, t) \doteq \sigma'(V_x(x, t))$, it can easily be checked that conditions (b.1), (b.2), and (b.3) are satisfied, so $\mathbb{T}$ is well defined. That $\mathbb{T}$ maps $\mathbb{X}$ into $\mathbb{X}$ for M sufficiently large and t_0 sufficiently small follows from (4.8).

Let $V, \bar{V} \in \mathbb{X}(M, t_0)$ and $U = \mathbb{T} V$, $\bar{U} = \mathbb{T} \bar{V}$; then $W \doteq U - \bar{U}$ satisfies

$$\mathcal{L} W \doteq W_t - \sigma'(V_x) W_{xx} - (\sigma'(V_x) - \sigma'(\bar{V}_x)) \bar{U}_{xx} = 0$$

and completely homogeneous initial-boundary conditions. From

$$\int_0^t \int_0^1 D_x^{\alpha_1} \mathcal{L} W \cdot D_x^{\alpha_1} W\,dx\,ds = 0, \qquad \alpha_1 = 0, 1,$$

we can show that

$$\int_0^1 W^2(x, t)\,dx \leqq k_1 t_0\, e^{k_2 t_0} \|V - \bar{V}\|_z^2,$$

and

$$\int_0^1 W_x^2(x, t)\, dx \leqq k_3 t_0 \| V - \bar{V} \|_z^2,$$

where

$$k_1 \doteq \sup_{\substack{\bar{Q} \\ 0 \leqq \theta \leqq 1}} |\sigma''(\theta V_x + (1-\theta)\bar{V}_x)\bar{U}_{xx}|^2,$$

$$k_2 \doteq 1 + \sigma_0^{-1} \sup_{\bar{Q}} |\sigma''(V_x) V_{xx}|^2,$$

$$k_3 \doteq \sigma_0^{-1} \sup_{\substack{\bar{Q} \\ 0 \leqq \theta \leqq 1}} |\sigma''(\theta V_x + (1-\theta)\bar{V}_x)\bar{U}_{xx}|.$$

Putting these together, we have

$$\| U - \bar{U} \|_z \leqq \kappa \| V - \bar{V} \|_z,$$

for $\kappa < 1$, provided t_0 is sufficiently small. □

We also include a sketch of the proof of existence and uniqueness for problem (P_1) in order to indicate precisely how the smoothness of $U(x, t)$ is intimately involved in obtaining a classical solution $U_1(x, t)$. This is recorded in the following theorem.

THEOREM 4. *Let σ, ϕ, and f satisfy the hypotheses of Theorem* 3 *and suppose*

$$(\psi.2) \qquad \psi \in H^6(0, 1), \qquad \psi = \psi'' = \psi^{iv} = 0 \quad \text{at } x = 0, 1.$$

Then the linear parabolic problem (P_1) has a unique classical solution $U_1(x, t) \in C^{2,1}(\bar{Q})$ such that

$$(r.7) \qquad U_{1_{tx}}, U_{1_{xxx}} \in C(\bar{Q}),$$

$$(r.8) \qquad U_{1_{tt}}, U_{1_{txx}}, U_{1_{ttx}}, U_{1_{ttt}}, U_{1_{xxxx}}, U_{1_{txxx}}, U_{1_{ttxx}}, U_{1_{xxxxx}}, U_{1_{txxxx}} \in L^\infty(L^2),$$

$$(r.9) \qquad U_{1_{tttx}}, U_{1_{ttxxx}}, U_{1_{xxxxxx}}, U_{1_{txxxxx}} \in L^2(Q).$$

Proof. It is convenient to consider the equation

$$\mathbb{L} U \doteq U_{1_t} - (c(x, t) U_{1_x})_x = 0,$$

where we assume that

$$(c.1) \qquad 0 < c_0 \leqq c(x, t),$$

$$(c.2) \qquad c_x = c_{xxx} = 0 \quad \text{at } x = 0, 1,$$

$$(c.3) \qquad c \in H^3(Q), \qquad D_x^{2\alpha_1 - 1} c \in H^{4-\alpha_1}(Q), \qquad \alpha_1 = 1, 2, 3.$$

We find that the Faedo-Galerkin approximations

$$W_m(x, t) = \sum_{k=1}^{m} C_{km}(t) \sin k\pi x$$

satisfy

$$(4.9) \qquad \langle D^\alpha \mathbb{L} W_m, D^\alpha W_m \rangle = 0$$

for $\alpha \in \mathbb{J} = \{\alpha = (\alpha_1, \alpha_2) | 0 \leqq \alpha_1 + 2 \max\{\frac{1}{2}, \alpha_2\} \leqq 6\}$, where $|\mathbb{J}| = 15$. We note that (c.2) is needed when $\alpha_1 = 3$ and 5. From (4.9) it follows that

$$\frac{1}{2}\int_0^1 |D^\alpha W_m|^2\,dx + \int_0^t \int_0^1 c(x,s)|D^\alpha W_{m_x}|^2\,dx\,ds$$
$$= \frac{1}{2}\int_0^1 |D^\alpha W_m(x,0)|^2\,dx - \int_0^t \int_0^1 D^\alpha W_{m_x} \sum_{\substack{0 \leqq \beta \leqq \alpha \\ 0 < |\beta|}} \binom{\alpha}{\beta} D^\beta c D^{\alpha-\beta} W_{m_x}\,dx\,ds$$

holds for each $\alpha \in \mathbb{J}$. Define

$$E^*(t) \doteq \frac{1}{2}\sum_{\alpha \in \mathbb{J}} \int_0^1 |D^\alpha W_m|^2\,dx;$$

then, with estimates similar to those in the Appendix, we arrive at the inequality

$$E^*(t) + \frac{1}{2}c_0 \int_0^t \int_0^1 \sum_{\alpha \in \mathbb{J}} |D^\alpha W_{m_x}|^2\,dx\,ds \leqq E^*(0) + \int_0^t \mathbb{K}(s)E^*(s)\,ds, \tag{4.10}$$

in which

$$\mathbb{K}(s) \doteq C\left\{\sum_{\alpha \in \mathbb{J}_1} \|D^\alpha c\|^2_{L^\infty(Q)} + \int_0^1 \sum_{\alpha \in \mathbb{J}_2} |D^\alpha c(x,s)|^2\,dx\right\}$$

where

$$\mathbb{J}_1 = \{\alpha = (\alpha_1, \alpha_2) | 1 \leqq \alpha_1 + 2\alpha_2 \leqq 3\}$$

and

$$\mathbb{J}_2 = \{\alpha = (\alpha_1, \alpha_2) | 4 \leqq \alpha_1 + 2\alpha_2 \leqq 6,\ \alpha \neq (6,0)\}.$$

It is important that the conditions imposed on the data of the problem enable us to bound $E^*(0)$ independently of m. Note that we may write

$$E^*(0) = \frac{1}{2}\sum_{\alpha_1=0}^{5} \int_0^1 |D_x^{\alpha_1} W_m(x,0)|^2\,dx + \frac{1}{2}\sum_{\substack{\alpha \in \mathbb{J} \\ \alpha_2 > 0}} \int_0^1 |D^\alpha W_m(x,0)|^2\,dx.$$

For the first sum, we obtain directly from the initial condition, by means of Bessel's inequality, the bound

$$\int_0^1 |D_x^{\alpha_1} W_m(x,0)|^2\,dx = \int_0^1 |\psi_m^{(\alpha_1)}(x)|^2\,dx \leqq \int_0^1 |\psi^{(\alpha_1)}(x)|^2\,dx$$

for $\alpha_1 = 0, 1, \cdots, 5$. For the second sum, we note that the orthogonality relations

$$\langle D^\alpha \mathbb{L} W_m, D^\alpha W_{m_t}\rangle = 0, \qquad \alpha \in \mathbb{J}_3,$$

where $\mathbb{J}_3 = \{\alpha = (\alpha_1, \alpha_2) | 0 \leqq \alpha_1 + 2\alpha_2 \leqq 4\}$, $|\mathbb{J}_3| = 9$, can be used to get

$$\int_0^1 |D^\alpha W_m(x,0)|^2\,dx \leqq C \int_0^1 \sum_{0 \leqq \beta \leqq \gamma} |D^\beta c(x,0) D^{\gamma-\beta} W_{m_x}(x,0)|^2\,dx,$$

where $\gamma = (\alpha_1 + 1, \alpha_2 - 1)$ for $\alpha \in \mathbb{J}$, $\alpha_2 > 0$, and C is a positive integer. This serves to bound each term of the second sum by initial data. It is clear that we need the traces

$$D^\alpha c(x,0) \in L^2(0,1), \qquad \alpha \in \{\alpha = (\alpha_1, \alpha_2) | 0 \leqq \alpha_1 + 2\alpha_2 \leqq 5\}.$$

Our assumption (c.3) is sufficient to guarantee that these traces exist. It is also at this point that we require $\psi \in H^6(0,1)$.

The proof now proceeds in a familiar way by applying Gronwall's lemma to (4.10). Uniqueness follows since the difference of two $C^{2,1}(\bar{Q})$ solutions, $Z \doteq U_1 - \hat{U}_1$, satisfies the identity

$$\frac{1}{2}\int_0^1 Z^2(x, t)\,dx + \int_0^t \int_0^1 c(x, s) Z_x^2(x, s)\,dx\,ds = 0.$$

The regularity of $U(x, t)$, the classical solution of (P_0), provided by Theorem 3, has been designed precisely so that $c(x, t) \doteq \sigma'(U_x(x, t))$ satisfies (c.1), (c.2), and (c.3). □

5. The main result. The analysis of the previous three sections permits some brevity here. In particular, when we form the problem for the remainder, analysis of existence and uniqueness is not needed since the remainder is a sum of functions whose regularity properties have been established. We have also investigated the order ε behavior of the derivatives of u so that here we can proceed with certain knowledge. All that is required at this point is the establishment of an a priori estimate for the remainder, allowing determination of its order ε behavior. Our main result then is contained in the following theorem.

THEOREM 5. *Let f, ϕ, ψ, and σ satisfy the same hypotheses as in Theorem* 4. *Denoting, as above, by Q the rectangular region $(0, 1)\times(0, t_0)$, let u be the solution of problem* (P_ε):

$$\begin{aligned} &\varepsilon^2 u_{tt} + u_t - \sigma'(u_x)u_{xx} = f(x, t), && (x, t) \in Q,\\ &u(x, 0; \varepsilon) = \phi(x), \qquad \varepsilon u_t(x, 0) = \psi(x), && x \in [0, 1],\\ &u(0, t; \varepsilon) = u(1, t; \varepsilon) = 0, && t \in [0, t_0], \end{aligned}$$

with the properties established in Theorem 2. *For the same t_0, let U and U_1, respectively, be the zeroth-order and first-order outer solutions given by Theorems* 3 *and* 4. *Let εV be the lowest-order initial-layer corrector given by* (4.1). *Then*

$$u(x, t; \varepsilon) = U(x, t) + \varepsilon U_1(x, t) + \varepsilon V(x, t/\varepsilon^2) + \mathcal{O}(\varepsilon^2) \quad \text{as } \varepsilon \to 0^+$$

uniformly on $\bar{Q}$. The interval $[0, t_0]$ may be small, but it is independent of ε in $(0, \varepsilon_0]$.

Proof. Let r denote the remainder term defined by

$$r \doteq u - [U + \varepsilon U_1 + \varepsilon V + \varepsilon^2 V_1],$$

where $\varepsilon^2 V_1$ is the corrector given by (4.2). For ease of notation, let

$$e \doteq U + \varepsilon U_1 + \varepsilon V + \varepsilon^2 V_1.$$

An easy computation shows that r satisfies the nonlinear hyperbolic equation

$$\varepsilon^2 r_{tt} + r_t - \sigma'(u_x)r_{xx} - b(r_x) = h(x, t; \varepsilon)$$

in the classical sense in $\bar{Q}$, where

$$b(r_x) \doteq [\sigma'(e_x + r_x) - \sigma'(e_x)]\, e_{xx},$$

and

$$\begin{aligned} h(x, t; \varepsilon) \doteq &-\varepsilon^2 U_{tt} - \varepsilon^3 U_{1_{tt}} + [\sigma'(e_x) - \sigma'(U_x)]U_{xx}\\ &-\varepsilon \frac{\partial}{\partial x}[\sigma'(U_x)U_{1_x}] + \sigma'(e_x)[\varepsilon U_{1_{xx}} + \varepsilon V_{xx} + \varepsilon^2 V_{1_{xx}}]. \end{aligned}$$

The remainder r also assumes the initial conditions

$$r(x, 0; \varepsilon) = -\varepsilon^2[\sigma'(\phi'(x))\phi''(x) + f(x, 0)]$$

and

$$\varepsilon r_t(x, 0; \varepsilon) = -\varepsilon^2[\sigma''(\phi'(x))\phi''(x)\psi'(x) + \sigma'(\phi'(x))\psi''(x)]$$

for $x \in [0, 1]$, as well as the homogeneous boundary conditions

$$r(0, t; \varepsilon) = r(1, t; \varepsilon) = 0, \qquad t \in [0, t_0].$$

By means of energy integrals, the following pointwise estimate can be established for such a function r

$$|r(x, t; \varepsilon)|^2 \leqq C\left\{\hat{E}(0) + \frac{1}{2}\int_0^{t_0}\int_0^1 h^2\, dx\, ds\right\}, \tag{5.1}$$

where the constant C is independent of ε in $(0, \varepsilon_0]$ and $\hat{E}$ is the energy

$$\hat{E}(t) \doteq \frac{1}{2}\int_0^1 [\varepsilon^2 r_t^2 + \sigma'(u_x) r_x^2]\, dx.$$

Assuming for the moment the validity of the bound (5.1), the theorem is completed by showing that the right-hand side is $\mathcal{O}(\varepsilon^4)$. With the initial conditions satisfied by r, this is obvious for $\hat{E}(0)$. The integral of the square of h over Q will also be $\mathcal{O}(\varepsilon^4)$ provided this holds for each of the terms whose sum comprises h. Thanks to the regularity of U and U_1 obtained in Theorems 3 and 4, this assertion clearly holds for the first two terms in h. A little more care is required for the remaining terms, which, by use of Taylor's theorem, may be written as the sum

$$\begin{aligned}
&[\varepsilon\sigma''(U_x)U_{xx}]V_x + [\varepsilon\sigma'(U_x)]V_{xx} + [\varepsilon^2\sigma''(U_x)U_{xx}]V_{1_x} + [\varepsilon^2\sigma'(e_x)]V_{1_{xx}} \\
&\qquad + \varepsilon^2[(U_{1_{xx}} + V_{xx})\sigma''(\theta e_x + (1-\theta)U_x)(U_{1_x} + V_x + \varepsilon V_{1_x})] \\
&\qquad + \varepsilon^2[\tfrac{1}{2}\sigma''(\theta e_x + (1-\theta)U_x)U_{xx}(U_{1_x} + V_x + \varepsilon V_{1_x})^2].
\end{aligned}$$

The last two terms here contribute $\mathcal{O}(\varepsilon^4)$ to the estimate, due to the regularity established in § 4 for U, U_1, V, and V_1. The remaining terms are each in the form of a product of an $\mathcal{O}(\varepsilon)$ factor and a corrector. When integration with respect to t is carried out in the integral over Q of the square of these terms, the factor $\exp\{-t/\varepsilon^2\}$ in the corrector gives rise to an additional ε^2, and thus, the result here is also seen to be $\mathcal{O}(\varepsilon^4)$.

It thus remains only to establish the uniform bound (5.1), which is done by estimating the right-hand side of the energy identity

$$\begin{aligned}
\hat{E}(t) + \int_0^t\int_0^1 r_t^2\, dx\, ds &= \hat{E}(0) + \int_0^t\int_0^1 r_t[b(r_x) + h]\, dx\, ds \\
&\quad + \int_0^t\int_0^1 \left\{\frac{1}{2} r_x^2 \frac{\partial}{\partial t}\sigma'(u_x) - r_x r_t \frac{\partial}{\partial x}\sigma'(u_x)\right\} dx\, ds.
\end{aligned} \tag{5.2}$$

We have by the arithmetic-geometric mean (A-G) inequality

$$\begin{aligned}
\int_0^t\int_0^1 |r_t b(r_x)|\, dx\, ds &\leqq \frac{1}{2\eta}\int_0^t\int_0^1 r_t^2\, dx\, ds \\
&\quad + \eta \sup_{\substack{\bar{Q} \\ 0<\theta<1}} \{|e_{xx}||\sigma''(\theta u_x + (1-\theta)e_x)|\}^2 \int_0^t\int_0^1 \frac{1}{2} r_x^2\, dx\, ds,
\end{aligned}$$

and then note that the last integral is bounded by

$$\sigma_0^{-1}\int_0^t \hat{E}(s)\,ds.$$

The integral of $r_t h$ is bounded by (A-G) with $\eta = 1$. Letting $a(x, t; \varepsilon) = \sigma'(u_x(x, t; \varepsilon))$, we have further that

$$\int_0^t\int_0^1 \left|\frac{1}{2}a_t r_x^2\right| dx\,ds \leqq \int_0^t \left(\sup_{0\leqq x\leqq 1}|a_t|\int_0^1 \frac{1}{2}r_x^2\,dx\right) ds$$

$$\leqq \sigma_0^{-1}\int_0^t \left\{1+\int_0^1 [a_t^2+a_{tx}^2]\,dx\right\} E(s)\,ds,$$

where Sobolev's inequality is used. The last term in (5.2) is treated similarly,

$$\int_0^t\int_0^1 |a_x r_x r_t|\,dx\,ds \leqq \frac{1}{2\eta}\int_0^t\int_0^1 r_t^2\,dx\,ds + \eta\int_0^t \left\{\left(\sup_{0\leqq x\leqq 1}|a_x|^2\right)\int_0^1 \frac{1}{2}r_x^2\,dx\right\} ds,$$

with the last term here bounded above by

$$\sigma_0^{-1}\eta\int_0^t \left\{\int_0^1 (a_x^2+a_{xx}^2)\,dx\right\}\hat{E}(s)\,ds.$$

Hence choosing $\eta = 2$ to cancel the second term on the left-hand side of (5.2) and letting

$$\mathcal{K}(s) = \frac{2}{\sigma_0}\left[\sup_{\substack{\bar{Q}\\ 0<\theta<1}} (|e_{xx}||\sigma''(\theta u_x + (1-\theta)\,e_x)|)^2\right.$$

$$\left. + \int_0^1 (a_x^2+a_{xx}^2)\,dx + \frac{1}{2}\left(1+\int_0^1 (a_t^2+a_{tx}^2)\,dx\right)\right],$$

we see that

$$\hat{E}(t) \leqq \hat{E}(0) + \frac{1}{2}\int_0^{t_0}\int_0^1 h^2\,dx\,ds + \int_0^t \mathcal{K}(s)\hat{E}(s)\,ds,$$

from which we obtain (5.1) by means of Gronwall's lemma and the Sobolev inequality. It is important to be aware that an ε-independent bound of the $L^1(0, t_0)$ norm of $\mathcal{K}(t)$ follows from the conclusions in Theorem 2 on the order ε behavior of the derivatives of the solution of the quasilinear hyperbolic equation. This completes the proof of the theorem. □

Appendix: Energy inequality. We wish first to establish (2.9):

$$\text{(A.1)}\quad E(t) + \frac{1}{4}\sum_{0\leqq|\alpha|\leqq 3}\varepsilon^{4\alpha_2}\int_0^t\int_0^1 |D^\alpha v_t|^2\,dx\,ds \leqq \left\{F(0) + \frac{1}{2}\|f_m\|^2_{H^3(Q)}\right\}\exp\int_0^t K(s)\,ds$$

for $v \equiv u_m$ given by (2.4). It will be clear that analogues of (A.1) are valid for higher-order α under appropriate conditions, but we operate under the hypotheses of Theorem 1, in which $|\alpha| \leqq 3$.

To prove (A.1) we start by considering, for any multi-index α, $0 \leqq |\alpha| \leqq 3$, the "energy" E_α defined by

$$E_\alpha(t) \equiv E_\alpha(t, \varepsilon; v) \doteq \frac{1}{2}\int_0^1 (\varepsilon^2|D^\alpha v_t|^2 + a(x, t; \varepsilon)|D^\alpha v_x|^2)\,dx,$$

where $D^\alpha \equiv D_x^{\alpha_1} D_t^{\alpha_2}$. Then our basic energy identity is

$$
\begin{aligned}
E_\alpha(t) + \int_0^t \int_0^1 |D^\alpha v_t|^2 \, dx \, ds &= E_\alpha(0) + \frac{1}{2} \int_0^t \int_0^1 a_t |D^\alpha v_x|^2 \, dx \, ds \\
&+ \int_0^t \int_0^1 D^\alpha v_t \left\{ \sum_{\substack{0 \leqq \beta \leqq \alpha \\ 0 < |\beta|}} \binom{\alpha}{\beta} D^\beta a D^{\alpha-\beta} v_{xx} - a_x D^\alpha v_x + D^\alpha f_m \right\} dx \, ds
\end{aligned} \tag{A.2}
$$

where f_m is the $L^2(0, 1)$ projection of f as defined in (2.8). The identity (A.2) is obtained from the $L^2(0, 1)$ orthogonality relation

$$\langle D^\alpha (L_\varepsilon v - f_m), D^\alpha v_t \rangle = 0, \qquad |\alpha| \leqq 3,$$

which follows directly from (2.5) by integration by parts. The vanishing of $a_x(x, t)$ at $x = 0$ and $x = 1$ is required in our hypotheses specifically so that this holds for $\alpha = (3, 0)$.

Multiplying (A.2) by $\varepsilon^{4\alpha_2}$ and summing over α, $0 \leqq |\alpha| \leqq 3$, yields

$$
\begin{aligned}
&F(t) + \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2} \int_0^t \int_0^1 |D^\alpha v_t|^2 \, dx \, ds \\
&\quad = F(0) + \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2} \int_0^t \int_0^1 D^\alpha v_t \\
&\qquad \times \left\{ \sum_{\substack{0 \leqq \beta \leqq \alpha \\ 0 < |\beta|}} \binom{\alpha}{\beta} D^\beta a D^{\alpha-\beta} v_{xx} - a_x D^\alpha v_x + D^\alpha f_m \right\} dx \, ds \\
&\qquad + \frac{1}{2} \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2} \int_0^t \int_0^1 a_t |D^\alpha v_x|^2 \, dx \, ds
\end{aligned} \tag{A.3}
$$

where, as in (2.10),

$$F(t) \doteq \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2} E_\alpha(t).$$

Defining, as in (2.9),

$$E(t) \doteq \frac{1}{2} \int_0^1 \left\{ \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2 + 2} |D^\alpha v_t|^2 + a(x, t; \varepsilon) \sum_{\alpha_1 = 0}^{3} |D_x^{\alpha_1} v_x|^2 \right\} dx,$$

we now derive the energy estimate

$$E(t) + \frac{1}{4} \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2} \int_0^t \int_0^1 |D^\alpha v_t|^2 \, dx \, ds \leqq F(0) + \frac{1}{2} \|f_m\|_{H^3(Q)}^2 + \int_0^t K(s) E(s) \, ds \tag{A.4}$$

from which (A.1) follows by using Gronwall's lemma.

There are several different estimates needed in order to bound the various terms in (A.3) to establish (A.4). Proper care must be exercised in associating powers of ε with derivatives of a and v so that only $\mathcal{O}(1)$ quantities appear. The term involving $D^\alpha f_m$ is bounded easily via the arithmetic-geometric mean inequality

$$\varepsilon^{4\alpha_2} \int_0^t \int_0^1 |D^\alpha v_t D^\alpha f_m| \, dx \, ds \leqq \frac{1}{2} \int_0^t \int_0^1 \{ \varepsilon^{4\alpha_2} |D^\alpha v_t|^2 + |D^\alpha f_m|^2 \} \, dx \, ds,$$

and hence the sum over α of such terms is bounded by

$$\frac{1}{2} \left\{ \|f_m\|_{H^3(Q)}^2 + \sum_{0 \leqq |\alpha| \leqq 3} \varepsilon^{4\alpha_2} \int_0^t \int_0^1 |D^\alpha v_t|^2 \, dx \, ds \right\} \tag{A.5}$$

where we have used, and will repeatedly use, the fact that $0<\varepsilon\leqq\varepsilon_0\leqq 1$. Similarly, by means of

$$\frac{1}{2}\varepsilon^{4\alpha_2}\int_0^t\int_0^1 |a_t||D^\alpha v_x|^2\,dx\,ds\leqq\frac{1}{2}\varepsilon^{4\alpha_2}\int_0^t \sup_{0\leqq x\leqq 1}|a_t|\left\{\int_0^1 |D^\alpha v_x|^2\,dx\right\}ds,$$

the sum involving the last term in (A.3) is bounded by

$$\mu\int_0^t \sup_{0\leqq x\leqq 1}|a_t(x,s)|E(s)\,ds, \tag{A.6}$$

where $\mu\doteq\max\{1, a_0^{-1}\}$. For the term involving $a_xD^\alpha v_x$, we use

$$\begin{aligned}&\varepsilon^{4\alpha_2}\int_0^t\int_0^1 |a_xD^\alpha v_tD^\alpha v_x|\,dx\,ds\\ &\leqq\frac{1}{2}\int_0^t \sup_{0\leqq x\leqq 1}|a_x(x,s)|\int_0^1\{\varepsilon^{4\alpha_2+2}|D^\alpha v_t|^2+\varepsilon^{4\alpha_2-2}|D^\alpha v_x|^2\}\,dx\,ds,\end{aligned} \tag{A.7}$$

when $\alpha_2>0$, while for $\alpha_2=0$, we have, for any $\eta>0$,

$$\int_0^t\int_0^1 |a_xD^\alpha v_tD^\alpha v_x|\,dx\,ds\leqq\frac{1}{2}\int_0^t\int_0^1\{\eta a_x^2|D^\alpha v_x|^2+\eta^{-1}|D^\alpha v_t|^2\}\,dx\,ds. \tag{A.8}$$

Thus the sum over α of such terms is bounded above by

$$\int_0^t\left\{\sup_{0\leqq x\leqq 1}|a_x(x,s)|+\eta\mu\sup_{0\leqq x\leqq 1}|a_x(x,s)|^2\right\}E(s)\,ds+\frac{1}{2\eta}\sum_{\alpha_1=0}^{3}\int_0^t\int_0^1|D_x^{\alpha_1}v_t|^2\,dx\,ds. \tag{A.9}$$

The remaining terms involve $D^\beta a$ and are a bit more sensitive. When $\beta=(1,0)$ the estimates (A.7) and (A.8) apply with $D^\alpha v_x$ replaced by $D^{\alpha-\beta}v_{xx}$. For $\beta=(0,1)$ we may use (A.7) if we replace a_x by a_t and $D^\alpha v_x$ by $D^{\alpha-\beta}v_{xx}$. For $|\beta|>1$, $\beta_2\neq 0$, we have

$$\begin{aligned}&\varepsilon^{4\alpha_2}\int_0^t\int_0^1 D^\alpha v_tD^\beta aD^{\alpha-\beta}v_{xx}|\,dx\,ds\\ &\leqq\int_0^t\left[\int_0^1\varepsilon^{4\beta_2-2}|D^\beta a|^2\,dx\right]^{1/2}\\ &\quad\times\left[\int_0^1\varepsilon^{4\alpha_2+2}|D^\alpha v_t|^2\varepsilon^{4(\alpha_2-\beta_2)}|D^{\alpha-\beta}v_{xx}|^2\,dx\right]^{1/2}ds\\ &\leqq\int_0^t\left[\int_0^1\varepsilon^{4\beta_2-2}|D^\beta a|^2\,dx\right]^{1/2}\\ &\quad\times\sup_{0\leqq x\leqq 1}|\varepsilon^{2(\alpha_2-\beta_2)}D^{\alpha-\beta}v_{xx}|\left[\int_0^1\varepsilon^{4\alpha_2+2}|D^\alpha v_t|^2\,dx\right]^{1/2}ds\\ &\leqq\frac{1}{2}\int_0^t\left[\int_0^1\varepsilon^{4\beta_2-2}|D^\beta a|^2\,dx\right]\left[\int_0^1\varepsilon^{4(\alpha_2-\beta_2)}|D^{\alpha-\beta}v_{xxx}|^2\,dx\right]ds\\ &\quad+\frac{1}{2}\int_0^t\int_0^1\varepsilon^{4\alpha_2+2}|D^\alpha v_t|^2\,dx\,ds.\end{aligned} \tag{A.10}$$

It should be mentioned that in (A.10) we have used the elementary Sobolev inequality

$$|D^\alpha v(x,t)|^2\leqq\int_0^1|D^\alpha v_x(\xi,t)|^2\,d\xi,$$

which is valid for the $C^4(\bar{Q})$ function $v = u_m$, in view of the fact that $D^\alpha v$ has a zero in $[0, 1]$ for each t.

For $|\beta| > 1$, $\alpha_2 = 0$, the same sort of estimation as in (A.10) yields

$$\int_0^t \int_0^1 |D^\alpha v_t D^\beta a D^{\alpha-\beta} v_{xx}|^2 \, dx \, ds$$

$$\leqq \frac{\eta}{2} \int_0^t \left[\int_0^1 |D_x^{\beta_1} a|^2 \, dx \right] \int_0^1 |D_x^{\alpha_1-\beta_1} v_{xxx}|^2 \, dx \, ds + \frac{1}{2\eta} \int_0^t \int_0^1 |D_x^{\alpha_1} v_t|^2 \, dx \, ds.$$

There is one additional case, when $\beta = (2, 0)$, $\alpha = (2, 1)$, for which, proceeding as in (A.10),

$$\varepsilon^4 \int_0^t \int_0^1 |D^\alpha v_t D^\beta a D^{\alpha-\beta} v_{xx}|^2 \, dx \, ds$$

$$\leqq \frac{1}{2} \int_0^t \left[\int_0^1 |D^\beta a|^2 \, dx \right] \int_0^1 \varepsilon^2 |D^{\alpha-\beta} v_{xxx}|^2 \, dx \, ds + \frac{1}{2} \int_0^t \int_0^1 \varepsilon^6 |D^\alpha v|^2 \, dx \, ds.$$

Merely collecting these estimates for the right-hand side and choosing η appropriately to obtain the coefficient $\frac{1}{4}$ for the double integral on the left-hand side, we see that there is a constant K_0, independent of ε, x, and t such that (A.1) holds with $K(s)$ given by (2.12). In simplifying the expression for K, we have used the elementary inequality

$$|a_x| \leqq \tfrac{1}{2}(1 + |a_x|^2)$$

as well as the Sobolev inequality

$$|\omega(x, t)|^2 \leqq C_0 \int_0^1 (|\omega(\xi, t)|^2 + |\omega_x(\xi, t)|^2) \, d\xi,$$

where C_0 is an absolute constant. With this, the proof of (A.1) is complete.

We now show that the right-hand side of (A.1) is bounded independently of m and ε. Consider first obtaining a uniform bound for the norms $\{\|f_m\|_{H^3(Q)}\}_{m=1}^\infty$. Note that $f \in H^3(Q)$ implies $D_t^{\alpha_2} f(\cdot, t) \in H^{3-\alpha_2}(0, 1)$, $\alpha_2 = 0, 1, 2, 3$, for almost all t in $(0, t_0)$ by Fubini's theorem. We have assumed that $f = f_{xx} = 0$ when $x = 0, 1$, and it follows that $f_t = f_{tt} = 0$ at $x = 0, 1$ as well. Since the family $\{\sin k\pi x\}_{k=1}^\infty$ is dense in $L^2(0, 1)$ and each of the spaces

$$S_k = \{v \in H^k(0, 1) | v = D_x^{2[k/3]} v = 0 \text{ at } x = 0, 1\},$$

$k = 1, 2, 3$, we find by means of Bessel's inequality that, for almost all t in $(0, t_0)$,

$$\int_0^1 |D^\alpha f_m|^2 \, dx \leqq \int_0^1 |D^\alpha f|^2 \, dx, \qquad |\alpha| \leqq 3. \tag{A.11}$$

Integrating over $(0, t_0)$ and adding, we find that $\|f_m\|_{H^3(Q)} \leqq \|f\|_{H^3(Q)}$. Next consider that

$$F(0) = \frac{1}{2} \sum_{|\alpha| \leqq 3} \varepsilon^{4\alpha_2} \int_0^1 \{\varepsilon^2 |D^\alpha u_{m_t}(x, 0)|^2 + a(x, 0; \varepsilon) |D^\alpha u_{m_x}(x, 0)^2\} \, dx$$

$$= \frac{1}{2} \sum_{\alpha_1=0}^{3} \int_0^1 \{\varepsilon^2 |D^\alpha u_{m_t}(x, 0)|^2 + a(x, 0; \varepsilon) |D^\alpha u_{m_x}(x, 0)|^2\} \, dx$$

$$+ \frac{1}{2} \sum_{\substack{|\alpha| \leqq 3 \\ \alpha_2 > 0}} \varepsilon^{4\alpha_2} \int_0^1 \{\varepsilon^2 |D^\alpha u_{m_t}(x, 0)|^2 + a(x, 0; \varepsilon) |D^\alpha u_{m_x}(x, 0)|^2\} \, dx.$$

From the initial conditions, for the first sum, we have

$$\frac{1}{2}\sum_{\alpha_1=0}^{3}\int_0^1 \{\varepsilon^2|D^\alpha u_{m_t}(x,0)|^2+a(x,0;\varepsilon)|D^\alpha u_{m_x}(x,0)|^2\}\,dx$$

$$\leqq\frac{1}{2}\sum_{\alpha_1=0}^{3}\int_0^1 \{|\psi^{(\alpha_1)}(x)|^2+a(x,0;\varepsilon)|\phi^{(\alpha_1+1)}(x)|^2\}\,dx$$

$$\leqq\frac{1}{2}\{\|\psi\|^2_{H^3(0,1)}+\|a(x,0;\varepsilon)\|_{L^\infty(0,1)}\|\phi\|^2_{H^4(0,1)}\}.$$

Also, the orthogonal relations

$$\langle D^\alpha(Lu_m-f_m),\, D^\alpha u_{m_{tt}}\rangle=0, \qquad |\alpha|\leqq 2,$$

can be used to get the bounds

$$\int_0^1 \varepsilon^{4\alpha_2+6}|D^\alpha u_{m_{tt}}(x,0)|^2\,dx$$

$$\leqq C\int_0^1\Bigg\{\sum_{0\leqq\beta\leqq\alpha}|(\varepsilon^{2\beta_2}D^\beta a(x,0))(\varepsilon^{2(\alpha_2-\beta_2)+1}D^{\alpha-\beta}u_{m_{xx}}(x,0))|^2$$

$$+\varepsilon^{4\alpha_2+2}|D^\alpha u_{m_t}(x,0)|^2+|D^\alpha f_m(x,0)|^2\Bigg\}\,dx,$$

for $|\alpha|\leqq 2$, in which we have been careful to place the available powers of ε so that only $\mathcal{O}(1)$ quantities appear on the right, in accordance with hypotheses (a.4) and (a.5), the initial conditions, and the results obtained for the α_2+1 case from the previous cases. Projections of the initial data (2.2) are easily bounded independently of m using Bessel's inequality, and an argument similar to that leading to (A.11) shows that

$$\int_0^1 |D^\alpha f_m(x,0)|^2\,dx\leqq\int_0^1 |D^\alpha f(x,0)|^2\,dx$$

holds for $|\alpha|\leqq 2$, with the traces existing at $t=0$ since $f\in H^3(Q)$. Note finally that the exponential factor in (A.1) is $\mathcal{O}(1)$ from the order ε behavior of the derivatives of $a(x,t;\varepsilon)$ assumed in (a.3) and (a.6).

REFERENCES

[1] A. BENAOUDA AND M. MADAUNE-TORT, *Singular perturbations for nonlinear hyperbolic-parabolic problems*, SIAM J. Math. Anal., 18 (1987), pp. 137-148.

[2] C. M. DAFERMOS AND J. A. NOHEL, *Energy methods for nonlinear hyperbolic Volterra integrodifferential equations*, Comm. Partial Differential Equations, 4 (1979), pp. 219-278.

[3] E. M. DE JAGER, *Singular perturbations of hyperbolic type*, Nieuw Arch. Wisk. (4), 23 (1975), pp. 145-171.

[4] E. M. DE JAGER AND R. GEEL, *Singular perturbations of hyperbolic type*, Analytical and Numerical Approaches to Asymptotic Problems in Analysis (Proc. Conf. Univ. Nijmegen, Nijmegen, 1980), pp. 57-71, North-Holland Math. Stud. 47, North-Holland, Amsterdam, 1981.

[5] B. F. ESHAM, *Asymptotics and an asymptotic Galerkin method for hyperbolic-parabolic singular perturbation problems*, SIAM J. Math. Anal., 18 (1987), pp. 762-776.

[6] B. F. ESHAM AND R. J. WEINACHT, *Hyperbolic-parabolic singular perturbations for scalar nonlinearities*, Appl. Anal., 29 (1988), pp. 19-44.

[7] R. GEEL, *Nonlinear initial value problems with singular perturbation of hyperbolic type*, Proc. Roy. Soc. Edinburgh Sect. A, 89 (1981), pp. 333-345.

[8] G. C. HSIAO AND R. J. WEINACHT, *A singularly perturbed Cauchy problem*, J. Math. Anal. Appl., 71 (1979), pp. 242-250.

[9] G. C. HSAIO AND R. J. WEINACHT, *Singular perturbations for a semilinear hyperbolic equation*, SIAM J. Math. Anal., 14 (1983), pp. 1168–1179.

[10] T. NISHIDA, *A note on the nonlinear vibrations of the elastic string*, Mem. Fac. Engrg. Kyoto Univ., 34 (1971), pp. 329–341.

[11] ———, *Nonlinear hyperbolic equations and related topics in fluid dynamics*, Publications Mathématiques d'Orsay 78-02, Département de Mathématique, Université de Paris-Sud, Orsay, France, 1978.

[12] J. L. NOWINSKI, *Theory of thermoelasticity with applications*, Sijthoff and Noordhoff, Alphen Aan Den Rijn, the Netherlands, 1978.

[13] M. SLEMROD, *Global existence, uniqueness, and asymptotic stability of classical smooth solutions in one-dimensional non-linear thermoelasticity*, Arch. Rational Mech. Anal., 76 (1981), pp. 97–133.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1366-1387, November 1989

006

SEMIDISCRETE APPROXIMATIONS OF HYPERBOLIC BOUNDARY VALUE PROBLEMS WITH NONHOMOGENOUS DIRICHLET BOUNDARY CONDITIONS*

I. LASIECKA† AND J. SOKOLOWSKI‡

Abstract. Finite-element approximations of the wave equation with nonhomogenous and "nonsmooth" Dirichlet boundary data are considered. These approximations are based on a special variational regularization of the problem introduced by J. L. Lions. The convergence rates of the approximations with nonsmooth boundary data are derived.

Key words. semidiscrete approximations, hyperbolic boundary value problems, nonhomogenous and nonsmooth Dirichlet data

AMS(MOS) subject classification. 35L

1. Introduction. Let Ω be an open bounded domain in R^n with smooth boundary Γ. Let $A(x, \partial)$ be a second-order strongly elliptic operator of the form

$$A(x, \partial)u = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}u\right)$$

where $a_{ij} = a_{ji} \in C^{\infty}(\bar{\Omega})$ and

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqq \alpha \sum_{i=1}^{n} \xi_i^2, \quad \alpha > 0 \quad \forall \xi_i \in R, \quad \forall x \in \Omega.$$

Consider the following second-order scalar hyperbolic equation:

$$\begin{aligned} &\ddot{u}(x, t) = A(x, \partial)u(x, t) && \text{in } \Omega \times (0, T) \equiv Q, \\ &u(x, 0) = \dot{u}(x, 0) = 0 && \text{in } \Omega, \\ &u(\sigma, t) = g(\sigma, t) && \text{in } \Gamma \times (0, T) \equiv \Sigma. \end{aligned} \tag{1.1}$$

The main goal of this paper is as follows. Under minimal regularity assumptions imposed on the boundary term g, we introduce finite-element approximation of (1.1) and establish convergence and the rates of convergence of the algorithm in $L_2(\Omega)$ norms. Our motivation for studying approximations of second-order hyperbolic equations with nonsmooth boundary data comes from problems arising in numerical considerations related to a variety of boundary control problems where the solutions are definitely nonsmooth, for example, optimization problems with boundary controls, the time-optimal boundary control problem, and Riccati equations arising from boundary control problems. To construct and prove related convergence of numerical algorithms for these problems, a preliminary step is to establish appropriate approximation of problem (1.1) with nonsmooth boundary data g—say, $g \in L_2(\Sigma)$ or $g \in H^1[0, T; H^{-1/2}(\Gamma)]$. To the authors' knowledge, the literature on finite-element methods for the second-order hyperbolic equation with Dirichlet boundary conditions deals only with nonhomogenous boundary data, which are smooth. This is not surprising, also in view of the fact that the maximal regularity of problem (1.1) with nonhomogenous boundary data $mL_2(\Sigma)$ has been established only recently (see [LT1],

* Received by the editors June 16, 1986; accepted for publication (in revised form) January 27, 1989.

† Mathematics Department, University of Florida, Gainesville, Florida 32611. The research of this author was partially supported by National Science Foundation grant DMS-8301168 and Air Force Office of Scientific Research grant AFOSR-84-0365.

‡ Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. This work was completed while the author was visiting the Mathematics Department of the University of Florida, Gainesville, Florida.

[L2], [LLT]). The presence of the nonhomogenous and nonsmooth Dirichlet boundary data is responsible for two immediate difficulties: (i) Dirichlet problem (1.1) does not admit a natural variational formulation that could then be taken as a basis for a numerical approximation. (ii) Low regularity of the boundary data g (hence of the solution) rules out the usual technique for proving stability and convergence of the numerical scheme based on $H^1(\Omega)\times L_2(\Omega)$ energy estimates. While the first difficulty can be handled by selecting an appropriate approximation of the elliptic operator that would take into account the nonhomogenous terms on the boundary (see, for example, [B2], [B7], [N1], [S1]), the second difficulty becomes crucial in relation to the derivation of stability estimates for the sought-after numerical algorithm. Let us elaborate more on this point. A standard finite-element approximation approach in the hyperbolic (as well as the parabolic) case is to define a semidiscrete algorithm by taking an appropriate space-approximation of the underlined elliptic operator. The estimates on the rate of convergence—which, of course, depend on the smoothness of the solutions—can be obtained by taking the difference of the two solutions and by using results on elliptic approximations. It is known, however [R1], that even if the elliptic approximations yield the optimal rates of convergence, nevertheless, the rates for hyperbolic problems are nonoptimal because they require one extra time-derivative of the solution. Since we cannot obtain optimal *convergence rates*, we would at least like to obtain *convergence* of the numerical algorithm in the "right topologies," i.e., where the maximal regularity of the map $g\to u$ takes place. To accomplish this we need to establish stability estimates for numerical schemes in precisely the same topologies (in fact, for the homogenous boundary data, this can be done by using the $H^1(\Omega)\times L_2(\Omega)$ energy methods mentioned earlier). This issue, however, raises another question: What is the maximal regularity of the map $g\to u$. As we have noted, this seemingly innocent question has only recently been answered in an optimal way (see [LT1], [L2], [LLT]). In the references above it is shown, in particular, that the map $g\to u$ is bounded from

$$L_2(\Sigma)\to C[0, T; L_2(\Omega)], \tag{1.2}$$

or more generally,

$$H^{s,s}(\Sigma)\to H^{s,s}(Q)\cap C[0, T; H^s(\Omega)], \qquad s\geqq 0 \tag{1.3}$$

where in (1.3) we must assume that for $s>\frac{1}{2}$, g satisfies some appropriate compatibility conditions at the origin. Results (1.2) and (1.3) improve by one-half derivative the previous results on regularity of solutions to (1.1) given in [LM]. Equipped with maximal regularity results for the original problem, we now wish to devise a numerical algorithm that can provide (i) the best possible rates of convergence (here we are resigned to "loosening" one derivative), and (ii) stability estimates reconstructing as much as possible the regularity properties of the original solution. Since our prime interest is to consider nonsmooth boundary data, it is precisely the second point mentioned above that limits our choices of elliptic approximations. The reason for this is twofold. First, the available elliptic estimates deal with more regular (in-space) boundary data—typically $g\in H^p(\Gamma)$; $p\geqq\frac{3}{2}$ (see [B2], [B7], [N1], [S1]). Second, standard techniques of proofs based on H^1-coercivity of elliptic problems are not applicable because we consider boundary data that do not yield $H^1(\Omega)$ solutions. Thus the sought-after elliptic approximation should allow for the treatment of nonsmooth boundary data g and, moreover, should be suitable for yielding hyperbolic estimates in lower norms. What we propose here is based on the idea originally introduced by Lions in [L1], where the original Dirichlet problem is "approximated" by the following sequence of problems with natural variational boundary conditions. For every $\varepsilon>0$,

parameter tending to zero, let u_ε be the solution of

$$\begin{aligned} &\ddot{u}_\varepsilon(x,t) = A(x,\partial)u_\varepsilon(x,t) && \text{in } Q,\\ &u_\varepsilon(x,0) = \dot{u}_\varepsilon(x,0) = 0 && \text{in } \Omega,\\ &\varepsilon\frac{\partial u_\varepsilon(\sigma,t)}{\partial\eta} + \beta u_\varepsilon(\sigma,t) = \beta g(\sigma,t) && \text{in } \Sigma \end{aligned} \tag{1.4}$$

where $\partial/\partial\eta$ stands for the co-normal derivative with respect to the operator A and where β is a second-order strongly elliptic self-adjoint operator defined on Γ (we can take $\beta = -\Delta_\Gamma + I$, where Δ_Γ is Laplace's Beltrami operator). This implies, in particular,

$$\langle \beta u, v\rangle \leqq C|u|_{H^1(\Gamma)}|v|_{H^1(\Gamma)}, \tag{1.4a}$$

$$\langle \beta u, u\rangle \geqq C_0|u|^2_{H^1(\Gamma)} \tag{1.4b}$$

where $c_1 c_0 > 0$ and $\beta^{-1}: H^s(\Gamma) \to H^{s+2}(\Gamma)$ exists and it is bounded.

Note that if we take $\beta \equiv I$ in (1.4), then the projection of (1.4) onto finite-dimensional subspaces of $H^1(\Omega)$ is a hyperbolic counterpart to the Penalty Method introduced by Babuška [B2] for elliptic problems. However, with $\beta = I$ in (1.4), the solution $u_\varepsilon(t)$ is *not bounded* in $L_2(\Omega)$ (uniformly with respect to the parameter $\varepsilon > 0$) by $|g|_{L_2(\Sigma)}$. We recall that the same bound holds true for the original solution (1.1) (see (1.21)). This shows that (1.4) with $\beta = I$ is *not a good* "approximation" of the original hyperbolic problem because it does not reconstruct the regularity properties of the original solutions. The presence of the Laplace Beltrami operator on the boundary forces stronger convergence of the traces of u_ε which, in turn, is necessary to obtain the appropriate stability of the solution (see [LS1], [L1]). In fact, it is shown in [LS1] (see also [L1]) that the following convergence result takes place.

THEOREM 1.1 [LS1].

(i) *For any* $g \in L_2(\Sigma)$

$$\|u_\varepsilon - u\|_{C[0,T;L_2(\Omega)]} \to 0,$$

$$\|u_\varepsilon\|_{C[0,T;L_2(\Omega)]} \leqq C|g|_{L_2(\Sigma)};$$

(ii) $$\|u_\varepsilon - u\|_{C[0,T;L_2(\Omega)]} \leqq C\varepsilon\|g\|_{H^{1,1}(\Sigma)};$$

(iii) $$\|u_\varepsilon - u\|_{C[0,T;H^1(\Omega)]} \leqq C\varepsilon\|g\|_{H^{2,2}(\Sigma)};$$

(iv) $$|u_\varepsilon|_\Gamma - g|_{L_2[0,T;H^2(\Gamma)]} \leqq C\varepsilon|g|_{H^{1,1}(\Sigma)}.$$

In defining a semidiscrete approximation of the original problem (1.1), a natural idea is to "project" the variational form of (1.4) onto the finite-dimensional subspaces. To this end, let h be the parameter of discretization tending to zero and let V_h stand for the approximating space of $H^1(\Omega)$ with the usual approximation properties (to be specified later) and such that $\tilde{V}_h \equiv V_{h|\Gamma} \subset H^1(\Gamma)$. As an approximation of $u_\varepsilon(t)$ (solution to (1.4)) we take $u_{h,\varepsilon}(t) \in V_h$ such that

$$\begin{aligned} &(\ddot{u}_{h,\varepsilon}(t), \phi_h)_\Omega + a(u_{h,\varepsilon}(t), \phi_h) + \frac{1}{\varepsilon}\langle \beta u_{h,\varepsilon}(t), \phi_h\rangle_\Gamma = \frac{1}{\varepsilon}\langle \tilde{P}_h g, \beta\phi_h\rangle_\Gamma,\\ &\phi_h \in V_h,\\ &u_{h,\varepsilon}(0) = \dot{u}_{h,\varepsilon}(0) = 0 \end{aligned} \tag{1.5}$$

where $a(u, v)$ is the bilinear form associated with $A(x, \partial)$, i.e.,

$$\begin{aligned} a(u,v) &= \sum_{i_j=1}^{n}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}, \frac{\partial v}{\partial x_i}\right)_\Omega\\ &= \sum_{i,j=1}^{n}\int_\Omega a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_j}\,dx \end{aligned}$$

and $\tilde{P}_h$ is the orthogonal projection form $L_2(\Gamma)$ onto $\tilde{V}_h$. Later we will use (1.5) with $\varepsilon = \varepsilon(h) = h^\gamma$ for some $\gamma > 0$. The corresponding solution will be denoted by u_h.

Remark 1.1. Note that the procedure described above (i) is well defined directly on $g \in L_2(\Sigma)$, and (ii) gives V_h as subspaces of $H^1(\Omega)$ that are not required to satisfy boundary conditions.

The main goal of this paper is to establish stability and the rates of convergence of the approximation $u_h(t)$ to the original solution $u(t)$. In fact, our main results in the case when V_h consists of piecewise linear functions (see Corollaries 3.1 and 3.2) establish, in particular, that with $\varepsilon = \varepsilon(h) = h$ in (1.5) we have

(1.6) (convergence)

(i) $\|u - u_h\|_{C[0,T;L_2(\Omega)]} \leqq Ch[|\dot{g}|_{H^1(\Sigma)} + |g|_{L_2[0,T;H^1(\Gamma)]}]$,

(ii) $\|u - u_h\|_{C[0,T;H^{1/2-\rho}(\Omega)]} \leqq Ch[|\dot{g}|_{H^{3/2}(\Sigma)} + |g|_{L_2[0,T;H^{3/2}(\Gamma)]}]$

where $\rho > 0$ is arbitrarily small;

(1.7) (stability)

(i) $\|u - u_h\|_{C[0,T;L_2(\Omega)]} \leqq C[|g|_{H^1[0,T;H^{-1/2+\rho}(\Gamma)]} + |g|_{L^2(\Sigma)}]$,

(ii) $\|u - u_h\|_{C[0,T;H^{1/2-\rho}(\Omega)]} \leqq C[|g|_{H^1[0,T;H^\rho(\Gamma)]} + |g|_{H^{1/2-\rho}(\Sigma)}]$

where C stands for a generic constant independent on $h > 0$ and g. For boundary data that display more regularity properties and satisfy the appropriate compatibility conditions, higher-order rates of convergence are given in Corollary 3.2.

Remark 1.2. Note that—in view of the optimal convergence results for the wave equation with homogenous boundary conditions, where an extra derivative in the solution is necessary (see [R1] and also [B1], [D1], [B3]) and optimal regularity of the solutions to Dirichlet problems (see 1.2)—estimates (1.6) are optimal. Although the stability results in (1.7) improve "almost" by one-half derivative the stability estimates implied by the convergence result in (1.6), they are still nonoptimal with respect to the sharp regularity of the solution u. In fact, $g \in L_2(\Sigma)$ will produce the solution $u \in C[0T; L_2(\Omega)]$ (see [L1], [LT2], [LLT]); thus we would expect the stability estimate (1.7)(i) to hold for any $g \in L_2(\Sigma)$ (instead of $\dot{g} \in L_2[0, T; H^{-1/2+\rho}(\Gamma)]$).

Remark 1.3. In the analysis of the approximation error, a crucial role is played by the very special behavior of the traces of hyperbolic solutions (see § 2). In fact, the solutions to wave equations are shown [LLT] to have better regularity on the boundary than interior regularity and the trace theorem would imply. This fact will be used in an essential way in the process of proving (1.6) and (1.7).

Remark 1.4. The results of numerical computations performed with linear splines on two-dimensional domain are given in [LS2]. In fact, the results of [LS2] confirm our theoretical findings presented in this paper.

The outline of the paper is as follows. In § 2 we discuss the properties and regularity of the continuous solution $u(t)$ as well as those of the regularized solution $u_\varepsilon(t)$. In § 3, we define semidiscrete approximating subspaces and approximations of (1.1). The main results of the paper, Theorems 3.1 and 3.2 and Corollaries 3.1 and 3.2, are stated at the end of § 3. The proofs of these results are relegated to § 5, while § 4 is devoted to a number of technical lemmas needed for those proofs.

The following notation will be used in the paper. $(\cdot,)$, (respectively, $\|\cdot\|$) denotes the usual $L_2(\Omega)$ inner product (respectively, the norm in $L_2(\Omega)$). $\langle \cdot, \cdot \rangle$ (respectively, $|\cdot|$) denotes the $L_2(\Gamma)$ inner product (respectively, the norm in $L_2(\Gamma)$). $H^s(\Omega)$, $H^{r,s}(Q)$, for $r, s > 0$, are the usual Sobolev spaces defined as in [LM]; if $r = s$ we use $H^r(Q) \equiv H^{r,r}(Q)$, $H^{-s} = (H^s)'$, $s \geqq 0$, where X' stands for the dual (pivotal) space to X. $\mathscr{L}(X \to Y)$ denotes the space of linear transformations from X to Y. $L_p[0, T; X]$, $1 \leqq p \leqq \infty$, denotes the space of $u(t) \in X$ such that $L_p[0, T]$ norm of $\|u(t)\|_x$ is well

defined. A^α, $0 \leqq \alpha \leqq 1$, stands for fractional powers of operator A (A^α are well defined for positive and self-adjoint operators; see, e.g., [P1]).

2. Regularity properties of the continuous solutions and some background material. To give a proper foundation to our approximation results, we should first ask what is the optimal regularity of the solution u for a given boundary data g. Until recently, the available regularity results (given in [LM]) were far from optimal. Only a few years ago, the issue of optimal regularity was settled. Below we collect some of these results.

THEOREM 2.1 [L2], [LLT], [LT1], [LT2]. *With $g \in L_2(\Sigma)$, let u be the solution to* (1.1). *Then*

(a) $u \in C[0, T; L_2(\Omega)]$,

(b) $\dot{u} \in C[0, T; (H_0^1(\Omega))']$,

(c) $\partial u/\partial \eta \in H^{-1,-1}(\Sigma)$.

If $g \in H^{r,r}(\Sigma)$, $r > 0$ and g satisfies the appropriate compatibility conditions, then

(a') $u \in C[0, T; H^r(\Omega)]$,

(b') $\dot{u} \in C[0, T; H^{r-1}(\Omega)]$,

(c') $\partial u/\partial \eta \in H^{r-1,r-1}(\Sigma)$.

Remark 2.1. Note that the boundary regularity results given in Theorem 2.1(c) and (c') do not follow from the interior regularity of u. Solutions behave "better" on the boundary than should follow from the interior regularity (a), (a') and the trace theory.

Regularity properties of the solutions $u_\varepsilon(t)$ to problem (1.4) will play a crucial role in establishing the error estimates for $u_{h,\varepsilon} - u_\varepsilon$. These properties have been studied recently in [L1] and [LS]. Below we collect some of these results.

THEOREM 2.2 [L1], [LS]. *Let u_ε be the solution to* (1.4). *Then*

(a) $\|u_\varepsilon\|_{C[0,T;L_2(\Omega)]} \leqq^1 C|g|_{L_2(\Sigma)}$.

Assuming additionally that $g(0) = 0$, we have

(b) $\|u_\varepsilon\|_{H^{1,1}(Q)} + \|u_\varepsilon\|_{C[0,T;H^1(\Omega)]} \leqq C|g|_{H^{1,1}(\Sigma)}$,

(c) $|u_\varepsilon|_\Gamma|_{L_2[0,T;H^1(\Gamma)]} \leqq C|g|_{H^{1,1}(\Sigma)}$,

(d) $|\dot{u}_\varepsilon|_\Gamma|_{L_2(\Sigma)} + \left|\frac{\partial u_\varepsilon}{\partial \eta_A}\right|_{L_2(\Sigma)} \leqq C|g|_{H^{1,1}(\Sigma)}$.

More generally, if g satisfies the appropriate compatibility conditions guaranteeing that $u_\varepsilon \in C[0, T; H^s(\Omega)]$ for $s > 1$, then

(e) $\|u_\varepsilon\|_{H^{s,s}(Q)} + \|u_\varepsilon\|_{C[0,T;H^s(\Omega)]} \leqq C|g|_{H^{s,s}(\Sigma)}$,

(f) $|u_\varepsilon|_\Gamma|_{L_2[0,T;H^s(\Gamma)]} \leqq C|g|_{H^{s,s}(\Sigma)}$,

(g) $|\dot{u}_\varepsilon|_\Gamma|_{H^{s-1,s-1}(\Sigma)} + \left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{H^{s-1,s-1}(\Sigma)} \leqq C|g|_{H^{s,s}(\Sigma)}$

where the constant C is uniform with respect to $\varepsilon > 0$.

[1] From now on we assume that a generic constant C depends on neither $\varepsilon > 0$ nor $h > 0$.

Remark 2.2. Note that the regularity properties of $u_\varepsilon(t)$ recover, uniformly in the parameter $\varepsilon > 0$, the regularity properties of the solution $u(t)$ to (1.1). This fact, together with convergence results given by Theorem 2.1, shows that (1.4) is a "good" approximation of (1.1). As already mentioned, the "more natural" approach—scheme (1.4) with $\beta = I$—does not have the same properties.

Later we will also use the regularity properties at the solutions to the following "elliptic problem":

$$\begin{aligned} &A(x,\partial)v_\varepsilon = f \quad \text{in } \Omega, \\ &\varepsilon\frac{\partial v_\varepsilon}{\partial \eta} + \beta v_\varepsilon = 0 \quad \text{on } \Gamma. \end{aligned} \tag{2.1}$$

Define the operator $A_\varepsilon : L_2(\Omega) \to L_2(\Omega)$ by $A_\varepsilon u \equiv A(x,\partial)u$ on $\mathscr{D}(A_\varepsilon) = \{u \in L_2(\Omega)\colon A(x,\partial)u \in L_2(\Omega);\ \varepsilon(\partial u/\partial \nu) + \beta u = 0 \text{ on } \Gamma\}$.

With the definition above, (2.1) is equivalent to

$$A_\varepsilon v_\varepsilon = f. \tag{2.1'}$$

It can be easily shown that A_ε is self-adjoint on $L_2(\Omega)$. The bilinear form $a_\varepsilon(u, v)$ associated with A_ε is given by

$$a_\varepsilon(u, v) \equiv -(A_\varepsilon u, v) = a(u, v) + 1/\varepsilon\langle \beta u, v\rangle \quad \text{for all } u, v \in H^1(\Omega);\ u|_\Gamma, v|_\Gamma \in H^1(\Gamma). \tag{2.2}$$

The following regularity results for (2.1) have been established in [LS].

LEMMA 2.3 [LS]. *Let $v_\varepsilon = A_\varepsilon^{-1} f$, the solution of* (2.1). *Then*

$$\|v_\varepsilon\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}}|v_\varepsilon|_{H^1(\Gamma)} + \left|\frac{\partial v_\varepsilon}{\partial \eta}\right|_{H^{-1/2}(\Gamma)} + \frac{1}{\varepsilon}|\beta v_\varepsilon|_{H^{-1/2}(\Gamma)} \leqq C\|f\|_{H^{-1}(\Omega)}, \tag{2.3}$$

$$\|v_\varepsilon\|_{H^2(\Omega)} + \frac{1}{\sqrt{\varepsilon}}|v_\varepsilon|_{H^2(\Gamma)} \leqq C\|f\|_{L_2(\Omega)}, \tag{2.4}$$

$$\|v_\varepsilon\|_{H^s(\Omega)} + \frac{1}{\sqrt{\varepsilon}}|v_\varepsilon|_{H^s(\Gamma)} \leqq C\|f\|_{H^{s-2}(\Omega)}, \qquad 1 \leqq s \leqq 2. \tag{2.5}$$

Later, we will find it convenient to represent the solution $u(t)$ of (1.1) as well as the solution $u_\varepsilon(t)$ of (1.4) in semigroup form. To accomplish this, let us define the operator $A: L_2(\Omega) \to L_2(\Omega)$ given by

$$Au = A(x,\partial)u, \qquad u \in \mathscr{D}(A) \equiv H_0^1(\Omega) \cap H^2(\Omega).$$

It is well known that A is the generator of an analytic semigroup on $L_2(\Omega)$ and it generates cosine $C(t)$ and sine $S(t)$ operators in $L_2(\Omega)$ (see [F1]). Next we define the so-called "Dirichlet map":

$D: L_2(\Gamma) \to L_2(\Omega)$ by

$$\begin{cases} A(x,\partial)Dg = 0 & \text{in } \Omega, \\ Dg|_\Gamma = g & \text{in } \Gamma. \end{cases} \tag{2.6}$$

From elliptic theory [LM] we know that

$$D \in \mathscr{L}(H^s(\Gamma) \to H^{s+1/2}(\Omega)) \quad \text{for all real } s. \tag{2.7}$$

Using the definitions above, we are in a position to represent the solution $u(t)$ in the semigroup form as in [LT1]:

$$u(t) = A\int_0^t S(t-z)Dg(z)\,dz \equiv (Lg)(t). \tag{2.8}$$

Theorem 2.1 gives

$$L \in \mathscr{L}(L_2(\Sigma) \to C[0, T; L_2(\Omega)]). \tag{2.9}$$

Similarly, we will represent the solution $u_\varepsilon(t)$ of (1.4) via the semigroup formula. To accomplish this we introduce the map $N_\varepsilon : L_2(\Gamma) \to L_2(\Omega)$ defined by

$$\begin{aligned} &A(x, \partial)N_\varepsilon g = 0 && \text{in } \Omega, \\ &\varepsilon\frac{\partial}{\partial\eta}(N_\varepsilon g) + \beta(N_\varepsilon g) = \beta g && \text{on } \Gamma. \end{aligned} \tag{2.10}$$

It can be shown [LS] that

$$N_\varepsilon \in \mathscr{L}(L_2(\Gamma) \to H^{1/2}(\Omega)) \text{ with the norm independent on } \varepsilon > 0. \tag{2.11}$$

The following identities are simple consequences of the Green formula (see [LS]):

$$\begin{aligned} &N_\varepsilon^* A_\varepsilon u = \frac{\partial}{\partial\eta} u \quad \forall u \in \mathscr{D}(A_\varepsilon), \\ &N_\varepsilon^* A_\varepsilon u = \frac{1}{\varepsilon}\beta u \quad \forall u \in C^\infty(\Omega). \end{aligned} \tag{2.12}$$

Since A_ε is self-adjoint and the spectrum of A_ε is on the real negative axis, A_ε generates cosine $C_\varepsilon(t)$ and sine $S_\varepsilon(t)$ operators on $L_2(\Omega)$. Therefore, following the same arguments as in [LT1], we show that the solution $u_\varepsilon(t)$ of (1.4) can be written as

$$u_\varepsilon(t) = A_\varepsilon\int_0^t S_\varepsilon(t-z)N_\varepsilon g(z)\,dz \equiv (L_\varepsilon g)(t). \tag{2.13}$$

From Theorem 2.2(a) it follows that

$$L_\varepsilon \in \mathscr{L}(L_2(\Sigma) \to C[0, T; L_2(\Omega)]) \text{ with the norm } \textit{independent} \text{ of } \varepsilon > 0. \tag{2.14}$$

Considering L_ε as acting from $L_2(\Sigma)$ into $L_2(Q)$, we compute its adjoint $L_\varepsilon^* : L_2(Q) \to L_2(\Sigma)$:

$$\begin{aligned} (L_\varepsilon^* f)(t) &= N_\varepsilon^* A_\varepsilon \int_t^T S_\varepsilon(z-t)f(z)\,dz \quad \text{by (2.8)} \\ &= \frac{\partial}{\partial\eta}\int_t^T S_\varepsilon(z-t)f(z)\,dz. \end{aligned} \tag{2.15}$$

As a consequence of (2.14), we have

$$L_\varepsilon^* \in \mathscr{L}(L_1[0, T; L_2(\Omega)] \to L_2(\Sigma)) \text{ with the norm independent of } \varepsilon > 0. \tag{2.16}$$

The solution $u_\varepsilon(t)$ given by (2.13) (or equivalently by (1.4)) can also be represented as the solution of the following abstract ordinary differential equation problem:

$$\begin{aligned} &\ddot{u}_\varepsilon(t) = A_\varepsilon u_\varepsilon(t) - A_\varepsilon N_\varepsilon g(t) \quad \text{on } \mathscr{D}(A_\varepsilon)', \\ &u_\varepsilon(0) = \dot{u}_\varepsilon(0) = 0. \end{aligned} \tag{2.17}$$

Equations (2.17) together with (2.8) and (2.14) lead to the following variational formulation of problem (1.4) (see also [L1]):

(2.18) $(\ddot{u}_\varepsilon(t), \phi)+a(u_\varepsilon(t), \phi)+1/\varepsilon\langle\beta u_\varepsilon(t), \phi\rangle=1/\varepsilon\langle\beta g, \phi\rangle$ for all $\phi\in\mathcal{D}(A_\varepsilon^{1/2})\equiv\{\phi\in H^1(\Omega), \phi_\Gamma\in H^1(\Gamma)\}$.

Note that the semidiscrete scheme (1.5) can be obtained from (2.18) by restricting the test functions ϕ to lie in the finite-dimensional subspace V_h.

3. Approximating subspaces and semidiscrete approximations of (1.4) and (1.1), and statement of main results. Let $h>0$ be the parameter of discretization tending to zero. Let V_h be a family of the finite-dimensional subspaces of $H^1(\Omega)$, of order $r\geqq 2$ satisfying local and inverse assumptions (see [B1, p. 98]) in addition to the following properties:

(3.1) (a) $V_h|_\Gamma\subset H^1(\Gamma)$;

(b) $\forall u\in H^s(\Omega)$

$$\inf_{\phi_h\in V_h}[\|u-\phi_h\|+h\|u-\phi_h\|_{H^1(\Omega)}+h^{1/2}|u-\phi_h|+h^{3/2}|u-\phi_h|_{H^1(\Gamma)}]$$

$$\leqq Ch^s\|u\|_{H^s(\Omega)}, \qquad \tfrac{3}{2}\leqq s\leqq r;$$

(c) $\|u-P_hu\|_{H^l(\Omega)}\leqq Ch^{s-l}\|u\|_{H^s(\Omega)}$, $0\leqq s\leqq r$, $0\leqq l<\frac{3}{2}$, $s-l\geqq 0$

where P_h is the orthogonal projection in $L_2(\Omega)$ (with respect to $L_2(\Omega)$ as inner product) onto V_h.

(d) For any $\phi_h\in\tilde{V}_h\equiv V_h|_\Gamma$,[2] there exists $\phi_h\in V_h$ such that $\phi_h|_\Gamma=\tilde{\phi}_h$ and $\|\phi_h\|_{H^s(\Omega)}\leqq C|\tilde{\phi}_h|_{H^{s-1/2}(\Gamma)}$, $0<s\leqq 1$.

It is well known that properties are standard and are satisfied by piecewise polynomials defined on the uniform mesh. Property (3.1)(d) with $s=1$ has been shown to be true in [B4] for polynomials defined on triangles. Arguments similar to those in [B4] have been used in [LS2] to prove that the inequality in (3.1)(d) can be extended to negative norms (i.e., $0<s<1$).

Below we state our main results on stability and the rate of convergence of the solution $u_{h,\varepsilon}(t)$ to $u_\varepsilon(t)$ in $H^s(\Omega)$ topology $0\leqq s<\frac{1}{2}$.

THEOREM 3.1 (stability). *Let u_ε be the weak solution of problem* (1.4) *and let $u_{h,\varepsilon}$ be the approximate solution of problem* (1.5). *Then with $\rho>0$ arbitrarily small, we have*:

(i) $\|u_{h,\varepsilon}\|_{C[0,T;H^{(1/2)-\rho}(\Omega)]}\leqq C|g|_{H^1[0,T;H^\rho(\Gamma)]}$.

(ii) $$\|u_{h,\varepsilon}\|_{C[0,T;L_2(\Omega)]}\leqq C\left[1+\frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right]|g|_{H^1[0,T;H^{-(1/2)+\rho}(\Gamma)]}$$

where $\sigma\geqq 1$ and $r\geqq 1+(\sigma-1)/2\rho$.

(iii) *If in addition we assume that $V_h^0\subset V_h$,[3] then*

$$\|u_{h,\varepsilon}\|_{C[0,T;L_2(\Omega)]}\leqq C|g|_{H^1[0,T;H^{-1/2+\rho}(\Gamma)]}.$$

Here C is independent on h, ε, and g.

THEOREM 3.2 (convergence). *Let u_ε (respectively, $u_{h,\varepsilon}$) be the solution of* (1.4) (*respectively*, (1.5)). *Assume that g satisfies the appropriate compatibility conditions at*

[2] It is well known that $\tilde{V}_h\subset H^1(\Gamma)$ is an approximating subspace of $L_2(\Gamma)$ of the same order r as V_h (see [B2, Thm. 4.22]).

[3] V_h^0 stands for the subspace of $H_0^1(\Omega)\cap V_h$ with approximating properties (3.1).

the origin to guarantee that $\dot{u} \in C[0T; H^s(\Omega)]$. *Then for any* $\rho > 0$ *arbitrarily small,* $\delta \geqq 1$, $s \geqq 1$ *there exists a constant* C *independent of* h, ε, *and* g *such that*:

(i) $$\|u_\varepsilon - u_{h,\varepsilon}\|_{C[0,T;L_2(\Omega)]} \leqq Ch^{s-\rho}\left[1+\frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right]^2 [|\dot{g}|_{H^s(\Sigma)}+|g|_{L_2[0,T;H^s(\Gamma)]}]$$

where $r \geqq 1+(s-1)(\sigma-1)/\rho$.[4]

(ii) $$\|u_\varepsilon - u_{h,\varepsilon}\|_{C[0,T;H^{(1/2)-\rho}(\Omega)]} \leqq Ch^{s-(1/2)-2\rho} f(h)[|\dot{g}|_{H^s(\Sigma)}+|g|_{L_2[0,T;H^s(\Gamma)]}]$$

where

$$f(h) \equiv \left\{ \begin{array}{ll} h^\rho & \text{if } s < \frac{3}{2} \\ \left(1+\dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right) & \text{if } r \geqq 1+\dfrac{(s-1)(\sigma-1)}{2\rho} \end{array} \right\}.$$

If in addition we assume that $V_h^0 \subset V_h$, *then for* $1 \leqq s \leqq r-\frac{1}{2}$ *we have*:

(iii) $$\|u_\varepsilon - u_{h,\varepsilon}\|_{C[0,T;L_2(\Omega)]} \leqq Ch^s[|\dot{g}|_{H^s(\Sigma)}+|g|_{L_2[0,T;H^s(\Gamma)]}].$$

(iv) $$\|u_\varepsilon - u_{h,\varepsilon}\|_{C[0,T;H^{(1/2)-\rho}(\Omega)]} \leqq Ch^{s-1/2}[|\dot{g}|_{H^s(\Sigma)}+|g|_{L_2[0,T;H^s(\Gamma)]}].$$

Let us set in (1.5) $\varepsilon = \varepsilon(h) = h^\gamma$ for some $\gamma > 1$, and let us denote the corresponding solution $u_{h,\varepsilon(h)}$ by u_h. After combining the results of Theorems 1.1, 3.1, and 3.2 we obtain the following corollary.

COROLLARY 3.1 (stability).

(i) $$\|u_h - u\|_{C[0,T;H^{(1/2)-\rho}(\Omega)]} \leqq C[|g|_{H^1[0,T;H^\rho(\Gamma)]}+|g|_{H^{(1/2)-\rho}(\Sigma)}].$$

(ii) $$\|u_h - u\|_{C[0,T;L_2(\Omega)]} \leqq C[|g|_{H^1[0,T;H^{-1/2+\rho}(\Gamma)]}+|g|_{L_2(\Sigma)}]$$

where $r \geqq 1+(1+\gamma)/2\rho$.

(iii) *If* $V_h^0 \subset V_h$ *then*

$$\|u_h - u\|_{C[0,T;L_2(\Omega)]} \leqq C[|g|_{H^1[0,T;H^{-(1/2)+\rho}(\Gamma)]}+|g|_{L_2(\Sigma)}].$$

COROLLARY 3.2 (convergence). *Let* u (*respectively,* u_h) *be the solution of problem* (1.1) (*respectively,* (1.5) *with* $\varepsilon(h) = h^\gamma$ *for some* $\gamma > 0$). *Then for any* $\rho > 0$ *arbitrarily small,* $s \geqq 1$ *we have the following*:

(i) $$\|u - u_h\|_{C[0,T;L_2(\Omega)]} \leqq C[h^{s-\rho}+h^\gamma][|\dot{g}|_{H^s(\Sigma)}+|g|_{L_2[0,T;H^s(\Gamma)]}]$$

where $r \geqq 1+(s-1)(1+\gamma)/\rho$.

(ii) $$\|u - u_h\|_{C[0,T;H^{(1/2)-\rho}(\Omega)]} \leqq C[h^{s-(1/2)-\rho}+h^\gamma][|\dot{g}|_{H^s(\Sigma)}+|g|_{L_2[0,T;H^s(\Gamma)]}]$$

where $r \geqq 1+(s-1)(1+\gamma)/\rho$ *if* $s \geqq \frac{3}{2}$.

(iii) *If in addition we assume that* $V_h^0 \subset V_h$, *then* (i) *and* (ii) *hold for any* $1 \leqq s \leqq r-\frac{1}{2}$ *and* $\rho = 0$.

Remark 3.1. Note that the rates of convergence established in part (iii) (respectively, (i), (ii)) are optimal (respectively, quasi-optimal) in the following sense: they reconstruct the optimal regularity of the solution (compare Theorem 2.1) (modulo the usual loss of one derivative). The estimates of the error given in part (iii) under the additional assumption that $V_h^0 \subset V_h$ reconstruct also the best approximation properties of the underlined approximating subspaces. If condition $V_h^0 \subset V_h$ fails, then for $s > 1$ we need to use higher-order polynomials to obtain the quasi-optimal error reflecting the optimal regularity of the solutions.

[4] If $s = 1$, then we can take $\rho = 0$, $\delta > 1$ arbitrary, and $r \geqq 1$.

Remark 3.2. Stability estimates provided by Corollary 3.1 improve by one-half derivative the stability results "implied" by the convergence results. Nevertheless, the stability estimates are still nonoptimal, since we are loosening one-half derivative with respect to the optimal regularity of the solutions (see Theorem 2.1).

Remark 3.3. We can, of course, interpolate between the results of Corollaries 3.1 and 3.2. For example, interpolation between the $L_2(\Omega)$-estimates of Corollaries 3.1 and 3.2 applied with $s=\gamma=1$ yields

$$\|u-u_h\|_{C[0,T;L_2(\Omega)]} \leqq Ch^{(1-Q)}[|\dot{g}|_{H^{1-3/2Q+Q\rho,1-Q}(\Sigma)}+|g|_{L_2[0,T;H^{1-3/2Q+Q\rho}(\Gamma)]}]$$

where $\rho>0$ is arbitrarily small and $0\leqq Q\leqq 1$.

The rest of the paper is devoted to the proofs of the main results.

4. Lemmas needed for the proofs of Theorems 3.1 and 3.2. The proofs of Theorems 3.1 and 3.2 will follow through the sequence of lemmas.

LEMMA 4.1. *Let $y\in H^{1+s}(\Omega)$ and $y|_\Gamma\in H^{1+s}(\Gamma)$. Then there exists $\hat{y}_h\in V_h$ such that*

$$\|y-\hat{y}_h\|_{H^1(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|y-\hat{y}_h|_{H^1(\Gamma)} \leqq Ch^s\left[\|y\|_{H^{1+s}(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|y|_{H^{1+s}(\Gamma)}\right], \qquad 0\leqq s<\frac{1}{2}. \tag{4.1}$$

If in addition we assume that $V_h^0\subset V_h$, then (4.1) *holds for $0\leqq s\leqq r-\frac{3}{2}$.* (4.2)

Otherwise (4.3)

$$\|y-\hat{y}_h\|_{H^1(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|y-\hat{y}_h|_{H^1(\Gamma)} \leqq C(\rho)h^{s-\rho}\left[\frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}+1\right]\|y\|_{H^{s+1}(\Omega)}+\frac{C}{\sqrt{\varepsilon}}h^s|y|_{H^{1+s}(\Gamma)}$$

where $\rho\geqq 0$, $0\leqq s\leqq r-\frac{3}{2}$, $\sigma\geqq 1$; $r\geqq 1+s(\sigma-1)/2\rho$, and C does not depend on h, ε, y.

Proof. Define $y^*=Dy|_\Gamma$, where $D: L_2(\Gamma)\to L_2(\Omega)$ is given by (2.6). From (2.7) it follows that

$$\|y^*\|_{H^{1+\alpha+1/2}(\Omega)} \leqq C|y|_\Gamma|_{H^{1+\alpha}(\Gamma)} \quad \text{for all } \alpha\in R. \tag{4.4}$$

Let $z\equiv y-y^*$. Then $z|_\Gamma=0$ and from (4.4) (applied with $\alpha=s-\frac{1}{2}$) it follows

$$\|z\|_{H^{1+s}(\Omega)} \leqq C[\|y\|_{H^{1+s}(\Omega)}+|y_{|\Gamma}|_{H^{1/2+s}(\Gamma)}] \quad \text{for all real } s. \tag{4.5}$$

If $s<\frac{1}{2}$, by Theorem 4.2.2 of [B1], we can select $z_h\in V_h$ such that

$$\|z-z_h\|_{H^1(\Omega)} \leqq Ch^s\|z\|_{H^{1+s}(\Omega)} \quad \text{and} \quad z_h|_\Gamma=0. \tag{4.6a}$$

If $V_h^0\subset V_h$, by the approximation properties of V_h^0 we can take $z_h\in V_h^0$ such that (4.6a) holds for all $0\leqq s\leqq r-1$. Otherwise, by applying Theorem 4.4 of [B3], we can select $z_h\in V_h$ such that

$$\begin{aligned}&\|z-z_h\|_{H^1(\Omega)} \leqq C(\rho)h^{s-\rho}\|z\|_{H^{s+1}(\Omega)},\\ &|z_h|_{H^1(\Gamma)} \leqq C(\rho)h^{s-\rho+\sigma/2-1}\|z\|_{H^{s+1}(\Omega)}\end{aligned} \tag{4.6b}$$

where $\rho>0$ is arbitrary, $\sigma\geqq 1$, and $r\geqq 1+s(\sigma-1)/2\rho$.

Next define $\hat{y}_h\equiv z_h+y_h^*\in V_h$, where $y_h^*\equiv P_hy^*$. From (3.1c) applied with $l=1$ and for $0\leqq s\leqq r-1$ it follows that

$$\begin{aligned}\|\hat{y}_h-y\|_{H^1(\Omega)} &\leqq \|z-z_h\|_{H^1(\Omega)}+\|y^*-y_h^*\|_{H^1(\Omega)} \leqq \|z-z_h\|_{H^1(\Omega)}+Ch^s\|y^*\|_{H^{1+s}(\Omega)}\\ &\leqq \|z-z_h\|_{H^1(\Omega)}+Ch^s|y_{|\Gamma}|_{H^{s+(1/2)}(\Gamma)},\end{aligned} \tag{4.7}$$

$$\frac{1}{\sqrt{\varepsilon}}|\hat{y}_h-y|_{H^1(\Gamma)} \leqq \frac{1}{\sqrt{\varepsilon}}|y^*-y_h^*|_{H^1(\Gamma)}+\frac{1}{\sqrt{\varepsilon}}|z_h|_{H^1(\Gamma)}. \tag{4.8}$$

On the other hand, by (3.1c), inverse approximation properties, and the approximating properties of $\tilde{V}_h$ it follows that

$$|y^* - P_h y^*|_{H^1(\Gamma)} \leqq Ch^s \|y^*\|_{H^{s+3/2}(\Omega)}, \qquad 0 \leqq s \leqq r - \tfrac{3}{2}.$$

Thus, for $0 \leqq s \leqq r - \frac{3}{2}$

$$|y^* - y_h^*|_{H^1(\Gamma)} \leqq Ch^s \|y^*\|_{H^{s+3/2}(\Omega)} \leqq Ch^s [y_{|\Gamma}|_{H^{s+1}(\Gamma)}] \tag{4.9}$$

where in the last inequality we have used again (2.7).

After collecting the results given in (4.5), (4.6a), and (4.7)-(4.9) we arrive at the desired conclusion stated in (4.1) and (4.2). As for (4.3), this follows from (4.5), (4.6b), and (4.7)-(4.9). □

LEMMA 4.1′. *Let* $y \in H^{1+s}(\Omega)$, $\partial y/\partial \eta|_\Gamma \in H^{s-1}(\Gamma)$, *and*

$$\varepsilon \frac{\partial y}{\partial \eta} + \beta y = \beta g_h \quad \text{on } \Gamma \quad \text{where } g_h \in \tilde{V}_h.$$

Then there exists $\hat{y}_h \in V_h$ *such that*

$$\|y - \hat{y}_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y - \hat{y}_h|_{H^1(\Gamma)} \leqq Ch^s \left[\|y\|_{H^{1+s}(\Omega)} + |g_h|_{H^{s+1}(\Gamma)} + \sqrt{\varepsilon} \left| \frac{\partial y}{\partial \eta} \right|_{H^{s-1}(\Gamma)} \right], \qquad 0 \leqq s < \frac{1}{2}. \tag{4.10}$$

If in addition $V_h^0 \subset V_h$, *then* (4.10) *holds for* $0 \leqq s \leqq r - \frac{3}{2}$. (4.11)

Otherwise, (4.12)

$$\|y - \hat{y}_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y - \hat{y}_h|_{H^1(\Gamma)} \leqq C(\rho) h^{s-\rho} \left(1 + \frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right) \cdot \left[\|y\|_{H^{1+s}(\Omega)} + |g_h|_{H^{s+1}(\Gamma)} + \sqrt{\varepsilon} \left| \frac{\partial y}{\partial \nu} \right|_{H^{s-1}(\Gamma)} \right]$$

where $\rho > 0$, $0 \leqq s \leqq r - \frac{3}{2}$, $\delta \geqq 1$, *and* $r \geqq 1 + s(\sigma - 1)/2\rho$.

Proof. Since $g_h \in \tilde{V}_h$, there exists $\phi_h \in V_h$ such that $\phi_h|_\Gamma = g_h$ and

$$\|\phi_h\|_{H^{s+1}(\Omega)} \leqq C|g_h|_{H^{s+1}(\Gamma)}, \qquad 0 \leqq s \leqq r - 1. \tag{4.13}$$

Let $\hat{y}_h^1$ be an approximation of $y - \phi_h$ selected according to Lemma 4.1.

Define $\hat{y}_h \equiv \hat{y}_h^1 - \phi_h$. Then

$$\|y - \hat{y}_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y - \hat{y}_h|_{H^1(\Gamma)} = \|y - \phi_h - \hat{y}_h^1\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y - \phi_h - \hat{y}_h^1|_{H^1(\Gamma)}. \tag{4.14}$$

Since $(y - \phi_h)|_\Gamma = -\varepsilon \beta^{-1} (\partial y/\partial \eta)$, we obtain

$$\frac{1}{\sqrt{\varepsilon}} |y - \phi_h|_{H^{1+s}(\Gamma)} \leqq \sqrt{\varepsilon} \left| \beta^{-1} \frac{\partial y}{\partial \eta} \right|_{H^{1+s}(\Gamma)} \leqq C\sqrt{\varepsilon} \left| \frac{\partial y}{\partial \eta} \right|_{H^{s-1}(\Gamma)} \tag{4.15}$$

where in the last inequality we have used the regularity of β as stated in (1.4′). Now, the assessments (4.10)-(4.12) of Lemma (4.1′) follow directly from Lemma 4.1 applied to $y - \phi_h$ and from the regularity properties (4.13) and (4.15). □

Next let us introduce the operator $P_{h,\varepsilon}: \mathcal{D}(A_\varepsilon^{1/2}) \to V_h$ defined as the projection onto V_h with respect to the norm generated by the bilinear form $a_\varepsilon(u, v)$. More precisely, $y_h \equiv P_{h,\varepsilon} y$ if and only if

$$a_\varepsilon(y_h - y, \phi_h) = 0 \quad \text{for all } \phi_h \in V_h. \tag{4.16}$$

The following lemma gives the estimates for the rate of convergence of $P_{h,\varepsilon} y$ to y.

LEMMA 4.2. *Let $y \in H^{1+s}(\Omega)$. Then*

$$\|P_{h,\varepsilon} y - y\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |P_{h,\varepsilon} y - y|_{H^1(\Gamma)} \leqq Ch^s \left[\|y\|_{H^{s+1}(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y|_\Gamma|_{H^{1+s}(\Gamma)} \right], \tag{4.17a}$$

$$\|P_{h,\varepsilon} y - y\|_{H^{1-s}(\Omega)} \leqq Ch^{2s} \left[\|y\|_{H^{1+s}(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y|_\Gamma|_{H^{1+s}(\Gamma)} \right], \qquad 0 \leqq s < \frac{1}{2}, \tag{4.17b}$$

If $V_h^0 \subset V_h$ then (4.17a) *remains valid for all $0 \leqq s \leqq r - \frac{3}{2}$,* (4.18a)

$$\|P_{h,\varepsilon} y - y\| \leqq Ch^{s+1} \left[\|y\|_{H^{1+s}(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y|_\Gamma|_{H^{1+s}(\Gamma)} \right], \qquad 0 \leqq s \leqq r - \frac{3}{2}, \tag{4.18b}$$

in the general case where $r \geqq 1 + s(\sigma - 1)/2\rho$, $0 \leqq s \leqq r - \frac{3}{2}$, $\sigma > 1$, $\rho \geqq 0$ arbitrary, we have

$$\begin{aligned} &\|P_{h,\varepsilon} y - y\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |P_{h,\varepsilon} y - y|_{H^1(\Gamma)} \\ &\quad \leqq C(\rho) h^{s-\rho} \left[1 + \frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}} \right] \left[\|y\|_{H^{1+s}(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y|_{H^{s+1}(\Gamma)} \right], \end{aligned} \tag{4.19a}$$

$$\|P_{h,\varepsilon} y - y\| \leqq C(\rho) h^{s+1-2\rho} \left(1 + \frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}} \right)^2 \left[\|y\|_{H^{1+s}(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |y|_{H^{s+1}(\Gamma)} \right]. \tag{4.19b}$$

Proof. For a given $y \in H^{1+s}(\Omega)$ let $\hat{y}_h$ be the element in V_h chosen according to Lemma 4.1. Then with $y_h \equiv P_{h,\varepsilon} y$ we have

$$a_\varepsilon(y_h - \hat{y}_h, v_h) = a_\varepsilon(y - \hat{y}_h, v_h), \qquad v_h \in V_h.$$

Setting $v_h = y_h - \hat{y}_h$ yields

$$\|y_h - \hat{y}_h\|_{H^1(\Omega)}^2 + \frac{1}{\varepsilon} |\beta^{1/2}(y_h - \hat{y}_h)|^2 \leqq C \left[\|\hat{y}_h - y\|_{H^1(\Omega)}^2 + \frac{1}{\varepsilon} |\hat{y}_h - y|_{H^1(\Gamma)}^2 \right].$$

Statements (4.17a), (4.18a), and (4.19a) are now direct consequences of Lemma 4.1 and triangle inequality.

As for (4.17b), (4.18b), and (4.19b) we use the "duality argument." Indeed, let $p \in L_2(\Omega)$. Define v_ε as the solution of

$$A_\varepsilon v_\varepsilon = p. \tag{4.20}$$

With the notation above we have

$$(y - P_{h,\varepsilon} y, P) = (y - P_{h,\varepsilon} y, A_\varepsilon v_\varepsilon) = a_\varepsilon(y - P_{h,\varepsilon} y, v_\varepsilon) = a_\varepsilon(y - P_{h,\varepsilon} y, v_\varepsilon - \hat{v}_{h\varepsilon})$$

where in the last step we have used (4.16) with $\hat{v}_{h\varepsilon}$—an approximation of v_ε chosen

according to Lemma 4.1. Hence

$$|(y-P_{h,\varepsilon}y,\,p)| \leqq C\|y-P_{h,\varepsilon}y\|_{H^1(\Omega)}\|v_\varepsilon-\hat{v}_{h,\varepsilon}\|_{H^1(\Omega)}+\frac{1}{\varepsilon}|y-P_{h,\varepsilon}y|_{H^1(\Gamma)}|v_\varepsilon-\hat{v}_{h,\varepsilon}|_{H^1(\Gamma)}$$

$$\leqq Ch^s\left[\|y\|_{H^{1+s}(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|y|_\Gamma|_{H^{1+s}(\Gamma)}\right] \tag{4.21}$$

$$\cdot\begin{cases} h^s\left[\|v_\varepsilon\|_{H^{1+s}(\Omega)}+\dfrac{1}{\sqrt{\varepsilon}}|v_\varepsilon|_{H^{1+s}(\Gamma)}\right], & 0\leqq s<\dfrac{1}{2} \\ h\left[\|v_\varepsilon\|_{H^2(\Omega)}+\dfrac{1}{\sqrt{\varepsilon}}|v_\varepsilon|_{H^2(\Gamma)}\right], & 0\leqq s\leqq r-1 \quad \text{if } V_h^0\subset V_h \end{cases}$$

(by Lemma 2.3, (2.4), and (2.5))

$$\leqq Ch^s\left[\|y\|_{H^{1+s}(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|y|_\Gamma|\right]_{H^{1+s}(\Omega)}\cdot\begin{cases} h^s\|p\|_{H^{s-1}(\Omega)}, & 0\leqq s<\frac{1}{2} \\ h\|p\|0\leqq s\leqq r-\frac{3}{2} & \text{if } V_h^0\subset V_h \end{cases}$$

which completes the proof of (4.17b) and (4.18b).

As for (4.19b), we use (4.3) in Lemma 4.1 to obtain

$$|(y-P_{h,\varepsilon}y,\,p)| \leqq C(\rho)h^{(s-\rho)}\left(1+\frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right)\left[\|y\|_{H^{s+1}(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|y|_{H^{s+1}(\Gamma)}\right]$$

$$\cdot\, h^{1-\rho}\left(1+\frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right)\left[\|v_\varepsilon\|_{H^2(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|v_\varepsilon|_{H^2(\Gamma)}\right]$$

$$\leqq C(\rho)h^{s+1-2\rho}\left(1+\frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right)^2\|p\|\left[\|y\|_{H^{s+1}(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|y|_{H^{s+1}(\Gamma)}\right],$$

which completes the proof of Lemma 4.2.

LEMMA 4.2′. *Let $y\in H^{1+s}(\Omega)$ and $\varepsilon(\partial y/\partial\eta)+\beta y=\beta g_h$ on Γ with $g_h\in\tilde{V}_h$. Denote*

$$|||y|||_{s,\varepsilon}\equiv\|y\|_{H^{1+s}(\Omega)}+|g_h|_{H^{1+s}(\Gamma)}+\sqrt{\varepsilon}\left|\frac{\partial y}{\partial\eta}\right|_{H^{s-1}(\Gamma)}.$$

With the notation above we have

$$\|P_{h,\varepsilon}y-y\|_{H^1(\Omega)}+\frac{1}{\sqrt{\varepsilon}}|P_{h,\varepsilon}y-y|_{H^1(\Gamma)}\leqq Ch^s|||y|||_{s,\varepsilon}, \tag{4.22a}$$

$$\|P_{h,\varepsilon}y-y\|_{H^{1-s}(\Omega)}\leqq Ch^{2s}|||y|||_{s,\varepsilon} \quad \textit{if } 0\leqq s<\tfrac{1}{2}. \tag{4.22b}$$

If $V_h^0\subset V_h$ and $0\leqq s\leqq r-\frac{3}{2}$, then

(4.23a) (4.22a) *holds for all $0\leqq s\leqq r-\frac{3}{2}$,*

$$\|P_{h,\varepsilon}y-y\|\leqq Ch^{s+1}|||y|||_{s,\varepsilon}, \tag{4.23b}$$

$$\|P_{h,\varepsilon}y-y\|_{H^{1-s_1}(\Omega)}\leqq Ch^{s_1+s}|||y|||_{s,\varepsilon} \quad \textit{where } 0\leqq s_1\leqq\tfrac{1}{2}. \tag{4.23c}$$

If $\rho > 0$, $0 \leqq s \leqq r - \frac{3}{2}$, $\delta \geqq 1$, *and* $r \geqq 1 + s(\sigma - 1)/2\rho$, *then*

$$\|P_{h,\varepsilon} y - y\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |P_{h,\varepsilon} y - y|_{H^1(\Gamma)} \leqq C(\rho) h^{s-\rho} \left(1 + \frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right) |||y|||_{s,\varepsilon}, \tag{4.24a}$$

$$\|P_{h,\varepsilon} y - y\| \leqq C(\rho) h^{s+1-2\rho} \left(1 + \frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right)^2 |||y|||_{s,\varepsilon}, \tag{4.24b}$$

$$\|P_{h,\varepsilon} y - y\|_{H^{1-s_1}(\Omega)} \leqq C(\rho) h^{s_1+s-\rho} \left(1 + \frac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right) |||y|||_{s,\varepsilon} \tag{4.24c}$$

where $0 \leqq s_1 < \frac{1}{2}$.

The proof of Lemma 4.2′ follows the same conceptual arguments as those used for the proof of Lemma 4.2. The only difference is that to estimate the terms in (4.21) we use Lemma 4.1′ instead of Lemma 4.1.

The following statements are the corollaries of Lemma 4.2.

COROLLARY 4.3. *Let* $A_{h,\varepsilon} : V_h \to V_h$ *be defined as the Galerkin approximation of* A_ε, *i.e.*,

$$(A_{h,\varepsilon} u_h, v_h) = (A_\varepsilon u_h, V_h) = a_\varepsilon(u_h, v_h), \qquad u_h, v_h \in V_h.$$

For any $f_h \in V_h$ *we have*

$$\begin{aligned} &\|(A_\varepsilon^{-1} - A_{h,\varepsilon}^{-1}) f_h\|_{H^{1-s}(\Omega)} \leqq C h^{2s} \|f_h\|_{H^{s-1}(\Omega)}, \qquad 0 \leqq s < \tfrac{1}{2}, \\ &\|(A_\varepsilon^{-1} - A_{h,\varepsilon}^{-1}) f_h\|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |(A_\varepsilon^{-1} - A_{h,\varepsilon}^{-1}) f_h|_{H^1(\Gamma)} \\ &\leqq \begin{cases} C h \|f_h\| & \text{if } V_h^0 \subset V_h, \\ C(\rho) h^{1-\rho} \left[1 + \dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right] \|f_h\| & \text{where } r \geqq 1 + \dfrac{\sigma - 1}{2\rho}, \quad \sigma > 1, \quad \rho > 0. \end{cases} \end{aligned} \tag{4.25}$$

Indeed, Corollary 4.3 follows directly from Lemma 4.2 after we set

$$v_\varepsilon \equiv A_\varepsilon^{-1} f_h, \qquad v_{h,\varepsilon} \equiv A_{h,\varepsilon}^{-1} f_h,$$

noting that

$$a_\varepsilon(v_\varepsilon - v_{h,\varepsilon}, \phi_h) = 0 \quad \text{for all } \phi_h \in V_h,$$

and using the regularity of v_ε as stated in Lemma 2.3. □

From Corollary 4.3 we also obtain Lemma 4.4.

LEMMA 4.4.

$$\|A_{h,\varepsilon}^{-1} f_h\|_{H^{s+1}(\Omega)} \leqq C \|f_h\|_{H^{s-1}(\Omega)}, \qquad 0 \leqq s < \tfrac{1}{2}.$$

Proof. By using the inverse approximation property, (2.5), and (4.25), we obtain

$$\begin{aligned} \|A_{h,\varepsilon}^{-1} f_h\|_{h^{s+1}(\Omega)} &\leqq \|(A_{h,\varepsilon}^{-1} - P_h A_\varepsilon^{-1}) f_h\|_{H^{s+1}(\Omega)} + \|P_h A_\varepsilon^{-1} f_h\|_{H^{s+1}(\Omega)} \\ &\leqq C h^{-2s} \|(A_{h,\varepsilon}^{-1} - P_h A_\varepsilon^{-1}) f_h\|_{H^{1-s}(\Omega)} + \|A_\varepsilon^{-1} f_h\|_{H^{s+1}(\Omega)} \\ &\leqq C \|f_h\|_{H^{s-1}(\Omega)}. \end{aligned}$$

□

Remark. Note that the estimate of Lemma 4.4 is a discrete counterpart of the regularity result stated in (2.5).

Next we prove Lemma 4.5.

LEMMA 4.5.

$$|\tilde{P}_h\beta A_{h,\varepsilon}^{-1}f_h|_{H^{s-1/2}(\Gamma)} \leqq C\varepsilon \|f_h\|_{H^{s-1}(\Omega)}, \qquad 0 \leqq s < \tfrac{1}{2}, \tag{4.26}$$

$$|\tilde{P}_h\beta A_{h,\varepsilon}^{-1}f_h|_{H^{(1/2)-\rho}(\Gamma)} \leqq C\varepsilon \begin{cases} \|f_h\| & \text{if } V_h^0 \subset V_h, \\ \left(1+\dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right)\|f_h\| & \text{where } \rho>0 \text{ arbitrarily small,} \\ \delta \geqq 1 \quad \text{and} \quad r \geqq 1+\dfrac{\sigma-1}{2\rho}. \end{cases} \tag{4.27}$$

Proof of Lemma 4.5. Let $v_h \equiv A_{h,\varepsilon}^{-1}f_h$ or, equivalently, $A_{h,\varepsilon}v_h = f_h$. Then

$$\frac{1}{\varepsilon}\langle \beta v_h, \phi_h\rangle = -(f_h, \phi_h) - a(v_h, \phi_h) \quad \text{for all } \phi_h \in V_h, \tag{4.28}$$

or after using the Green formula

$$\frac{1}{\varepsilon}(\beta v_h, \phi_h) = -(f_h, \phi_h) + (A(x,\partial)v_h, \phi_h) - \left\langle \frac{\partial}{\partial \eta} v_h, \phi_h \right\rangle, \qquad \phi_h \in V_h. \tag{4.29}$$

Now we will use the property (3.1d). For any $\tilde{\phi}_h \in \tilde{V}_h$ we take $\phi_h \in V_h$ such that $\phi_h|_\Gamma = \tilde{\phi}_h$ and for $0 < \alpha \leqq 1$, $\|\phi_h\|_{H^\alpha(\Omega)} \leqq C|\tilde{\phi}_h|_{H^{\alpha-(1/2)}(\Gamma)}$. Thus we can write

$$\frac{1}{\varepsilon}\langle \tilde{P}_h\beta v_h, \tilde{\phi}_h\rangle = \frac{1}{\varepsilon}\langle \beta v_h, \tilde{\phi}_h\rangle = -(f_h, \phi_h) - a(v_h, \phi_h). \tag{4.30}$$

Hence for $0 \leqq s < \frac{1}{2}$

$$\frac{1}{\varepsilon}\langle \tilde{P}_h\beta v_h, \tilde{\phi}_h\rangle \leqq C[\|f_h\|_{H^{s-1}(\Omega)}\|\phi_h\|_{H^{1-s}(\Omega)} + \|v_h\|_{H^{1+s}(\Omega)}\|\phi_h\|_{H^{1-s}(\Omega)}]$$

(by Lemma 4.4 applied to the second term above)

$$\begin{aligned} &\leqq C[\|f_h\|_{H^{s-1}(\Omega)}\|\phi_h\|_{H^{1-s}(\Omega)}] \\ &\leqq C\|f_h\|_{H^{s-1}(\Omega)}|\tilde{\phi}_h|_{H^{(1/2)-s}(\Gamma)} \end{aligned} \tag{4.31}$$

where in the last step we used (3.1d). Since (4.31) holds for all $\tilde{\phi}_h \in H^{(1/2)-s}(\Gamma)$, (4.26) follows via duality.

As for (4.27) we write

$$\begin{aligned} \frac{1}{\varepsilon}|\tilde{P}_h\beta A_{h,\varepsilon}^{-1}f_h|_{H^{(1/2)-\rho}(\Gamma)} &\leqq \frac{1}{\varepsilon}|\tilde{P}_h\beta(A_{h,\varepsilon}^{-1} - A_\varepsilon^{-1})f_h|_{H^{(1/2)-\rho}(\Gamma)} \\ &\quad + \frac{1}{\varepsilon}|\beta A_\varepsilon^{-1}f_h|_{H^{(1/2)-\rho}(\Gamma)}. \end{aligned} \tag{4.32}$$

We will prove that

$$\frac{1}{\varepsilon}|\beta A_\varepsilon^{-1}f|_{H^{1/2}(\Gamma)} \leqq C\|f\|. \tag{4.33}$$

Indeed, let $v_\varepsilon \equiv A_\varepsilon^{-1} f$. Then

$$\frac{1}{\varepsilon}\langle \beta v_\varepsilon, \phi\rangle = -a(v_\varepsilon, \phi) + (f, \phi) = (A(x,\partial)v_\varepsilon, \phi) - \left\langle \frac{\partial v_\varepsilon}{\partial \eta}, \phi \right\rangle + (f, \phi), \quad \phi \in D(A_\varepsilon^{1/2}). \tag{4.34}$$

Let $\tilde{\phi} \in H^{-1/2}(\Gamma)$. By (2.7), $\phi \equiv D\tilde{\phi} \in L_2(\Omega)$ and from (4.34) we obtain

$$\frac{1}{\varepsilon}\langle \beta v_\varepsilon, \tilde{\phi}\rangle \leqq C\left[\|v_\varepsilon\|_{H^2(\Omega)} \|D\tilde{\phi}\| + \left|\frac{\partial v_\varepsilon}{\partial \eta}\right|_{H^{1/2}(\Gamma)} |\tilde{\phi}|_{H^{-1/2}(\Gamma)} + \|f\|\,\|\phi\| \right] \quad \text{for all } \tilde{\phi} \in H^{-1/2}(\Gamma). \tag{4.35}$$

Hence by virtue of (2.7) and (2.4) in Lemma 2.3,

$$\frac{1}{\varepsilon}\langle \beta v_\varepsilon, \tilde{\phi}\rangle \leqq C\|f\|\,|\tilde{\phi}|_{H^{-1/2}(\Gamma)}$$

which via duality proves (4.33).

To complete the proof of (4.27), in view of (4.33), it is enough to estimate $1/\varepsilon |\tilde{P}_h \beta (A_{h,\varepsilon}^{-1} - A_\varepsilon^{-1}) f_h|_{H^{1/2-\rho}(\Gamma)}$.

With $v_{h,\varepsilon} \equiv A_{h,\varepsilon}^{-1} f_h$ and $v_\varepsilon \equiv A_\varepsilon^{-1} f_h$ it is straightforward to show that

$$a(v_{h,\varepsilon} - v, \phi_h) + \frac{1}{\varepsilon}\langle \beta(v_{h,\varepsilon} - v), \phi_h\rangle = 0 \quad \text{for } \phi_h \in V_h.$$

Thus for any $\tilde{\phi}_h \in \tilde{V}_h$ we have

$$\frac{1}{\varepsilon}\langle \tilde{P}_h \beta(v_{h,\varepsilon} - v), \tilde{\phi}_h\rangle = -a(v_h - v, \phi_h) \tag{4.36}$$

where ϕ_h is selected according to (3.1d), i.e., $\phi_h|_\Gamma = \tilde{\phi}_h$ and

$$\|\phi_h\|_{H^\rho(\Omega)} \leqq C|\tilde{\phi}_h|_{H^{-1/2+\rho}(\Gamma)} \tag{4.37}$$

where $\rho > 0$ arbitrarily small.

From (4.36) and Corollary 4.3 we obtain

$$\frac{1}{\varepsilon}\langle \tilde{P}_h \beta(v_{h,\varepsilon} - v), \tilde{\phi}_h\rangle \leqq C\|v_{h,\varepsilon} - v\|_{H^1(\Omega)} \|\phi_h\|_{H^1(\Omega)} \leqq \begin{cases} C\|f_h\|\,\|\phi_h\| & \text{if } V_h^0 \subset V_h, \\ C\left(1 + \dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right)\|f_h\|\,\|\phi_h\|_{H^\rho(\Omega)} \end{cases} \tag{4.38}$$

where in the last inequality we have used the inverse approximation property. Combining (4.37) with (4.38) and using duality yields

$$\frac{1}{\varepsilon}|\tilde{P}_h \beta(v_{h,\varepsilon} - v)|_{H^{(1/2)-\rho}(\Gamma)} \leqq C \begin{cases} \|f_h\| & \text{if } V_h^0 \subset V_h, \\ \left(1 + \dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right)\|f_h\| & \text{otherwise.} \end{cases} \tag{4.39}$$

Formulas (4.39) and (4.33) inserted into (4.32) complete the proof of (4.27), and, hence, of the lemma. □

In the sequel, we will find it convenient to use semigroup representations of the solution to (1.5), $u_{h,\varepsilon}(t)$. To this end let $S_{h,\varepsilon}(t)$ (respectively, $C_{h,\varepsilon}(t)$): $V_h \to V_h$ be the sine (respectively, cosine) operators associated with $A_{h,\varepsilon}$, where $A_{h,\varepsilon}$ is defined in Corollary 4.3. This means

$$\begin{aligned} &\ddot{C}_{h,\varepsilon}(t)x_h = A_{h,\varepsilon}x_h, \qquad C_{h,\varepsilon}(0)x_h = x_h, \\ &\dot{C}_{h,\varepsilon}(0)x_h = 0, \qquad\quad \dot{C}_{h,\varepsilon}(t) = A_{h,\varepsilon}S_{h,\varepsilon}(t). \end{aligned} \tag{4.40}$$

Using the notation above it can be easily verified that with $g_h \equiv \tilde{P}_h g$, $u_{h,\varepsilon}(t)$, the solution to (1.5) can be represented as

$$u_{h,\varepsilon}(t) = \int_0^t S_{h,\varepsilon}(t-z)P_hA_\varepsilon N_\varepsilon g_h(z)\,dz \equiv (L_{h,\varepsilon}g)(t) \tag{4.41}$$

or in the differential form as

$$\begin{aligned} &\ddot{u}_{h,\varepsilon}(t) = A_h u_{h,\varepsilon}(t) - P_hA_\varepsilon N_\varepsilon g_h(t), \\ &u_{h,\varepsilon}(0) = \dot{u}_{h,\varepsilon}(0) = 0. \end{aligned} \tag{4.42}$$

Remark. Note that although $A_\varepsilon N_\varepsilon$ is unbounded (even more so $\mathcal{D}(A_\varepsilon N_\varepsilon) \equiv 0$), $P_hA_\varepsilon N_\varepsilon g_h$ is well defined for any $g \in L_2(\Sigma)$. Indeed, by (2.17), $(P_hA_\varepsilon N_\varepsilon g_h, \phi_h) = \langle g_h, N_\varepsilon^* A_\varepsilon \phi_h\rangle = 1/\varepsilon\langle g_h, \beta\phi_h\rangle$ for all $\phi_h \in V_h$.

The next lemma deals with the approximation properties of sine $S_{h,\varepsilon}(t)$ and cosine $C_{h,\varepsilon}(t)$ operators.

LEMMA 4.6.

$$\|C_{h,\varepsilon}v_h\|_{L^\infty[0,T;L_2(\Omega)]} + \|S_{h,\varepsilon}v_h\|_{L^\infty[0,T;H^1(\Omega)]} \leqq C\|v_h\|, \tag{4.43}$$

$$\|C_{h,\varepsilon}v_h\|_{L^\infty[0,T;\mathcal{D}(A_\varepsilon^{1/2})]} \leqq C\|v_h\|_{\mathcal{D}(A_\varepsilon^{1/2})}, \tag{4.44}$$

$$\|C_{h,\varepsilon}v_h\|_{L^\infty[0,T;\mathcal{D}(A_\varepsilon^{\alpha})]} \leqq C\|v_h\|_{\mathcal{D}(A_\varepsilon^{\alpha})}, \qquad 0 \leqq \alpha \leqq \tfrac{1}{2}. \tag{4.45}$$

Proof. Define $y_h(t) \equiv S_{h,\varepsilon}(t)v_h$ with $v_h \in V_h$. Then

$$\ddot{y}_h(t) = A_{h,\varepsilon}y_h(t), \quad y_h(0) = 0, \quad \dot{y}_h(0) = v_h,$$

or equivalently,

$$(\ddot{y}_h(t), \phi_h) + a(y_h(t), \phi_h) + \frac{1}{\varepsilon}\langle \beta y_h(t), \phi_h\rangle = 0, \qquad \phi_h \in V_h. \tag{4.46}$$

Setting $\phi_h = \dot{y}_h(t)$ yields

$$\frac{d}{dt}\left[\|\dot{y}_h(t)\|^2 + \|y_h(t)\|^2_{H^1(\Omega)} + \frac{1}{\varepsilon}|\beta^{1/2}y_h(t)|^2\right] = 0.$$

Hence

$$\|C_{h,\varepsilon}(t)v_h\|^2 + \|S_{h,\varepsilon}(t)v_h\|^2_{H^1(\Omega)} + \frac{1}{\varepsilon}|\beta^{1/2}S_{h,\varepsilon}(t)v_h|^2 \leqq C\|v_h\|,$$

which gives (4.43).

To prove (4.44) we take $y_h(t) \equiv C_{h,\varepsilon}(t)v_h$. The same argument as above yields

$$\left[\|A_h S_{h,\varepsilon}(t)v_h\|^2 + \|C_{h,\varepsilon}(t)v_h\|^2_{H^1(\Omega)} + \frac{1}{\varepsilon}|\beta^{1/2}C_{h,\varepsilon}(t)v_h|^2\right] \leqq C\left[\|v_h\|^2_{H^1(\Omega)} + \frac{1}{\varepsilon}|\beta^{1/2}v_h|^2\right]. \tag{4.47}$$

Formula (4.44) follows from (4.47) after we note that

$$\|x\|^2_{D(A_\varepsilon^{1/2})} \sim \|x\|^2_{H^1(\Omega)} + \frac{1}{\varepsilon}|\beta^{1/2}x|^2.$$

Formula (4.45) is the result of interpolation between (4.43) and (4.45). □

5. Proofs of Theorems 3.1 and 3.2.

Proof of Theorem 3.1. Using the semigroup representation of $u_{h,\varepsilon}(t)$ given by (4.41) and integrating (4.41) by parts, we obtain

$$\begin{aligned} u_{h,\varepsilon}(t) = (L_{h,\varepsilon}g)(t) &= \int_0^t \frac{d}{dz} C_{h,\varepsilon}(t-z)A_{h,\varepsilon}^{-1}P_hA_\varepsilon N_\varepsilon g_h(z)\,dz \\ &= A_{h,\varepsilon}^{-1}P_hA_\varepsilon N_\varepsilon g_h(t) - C_{h,\varepsilon}(t)A_{h,\varepsilon}^{-1}P_hA_\varepsilon N_\varepsilon g_h(0) \\ &\quad - \int_0^t C_{h,\varepsilon}(t-z)A_{h,\varepsilon}^{-1}P_hA_\varepsilon N_\varepsilon \dot{g}_h(z)\,dr \end{aligned} \tag{5.1}$$

where we have used the properties of $S_{h,\varepsilon}(t)$ and $C_{h,\varepsilon}(t)$, the operators given in (4.40).

From $g_h = \tilde{P}_h g \in H^1[0, T; H^{-1/2}(\Gamma)]$ it follows, in particular, that

$$|g_h|_{C[0,T;H^{-1/2+\rho}(\Gamma)]} \leqq C|g_h|_{H^1[0,T;H^{-1/2+\rho}(\Gamma)]} \leqq C|g|_{H^1[0,T;H^{-1/2+\rho}(\Gamma)]}. \tag{5.2}$$

Thus, in view of (5.2) and (4.43) in Lemma 4.6, parts (ii) and (iii) of Theorem 3.1 will be proved as soon as we establish the following inequality:

$$\|A_{h,\varepsilon}^{-1}P_hA_\varepsilon N_\varepsilon \tilde{P}_h g\| \leqq C|g|_{\bar{H}^{1/2+\rho}(\Gamma)} \cdot \begin{cases} 1 & \text{if } V_h^0 \subset V_h, \\ \left(1 + \dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right), & r \geqq 1 + \dfrac{\sigma-1}{2\rho}. \end{cases} \tag{5.3}$$

To verify (5.3) we define

$$T_{h,\varepsilon}g \equiv A_{h,\varepsilon}^{-1}P_hA_\varepsilon N_\varepsilon \tilde{P}_h g \tag{5.4}$$

where $T_{h,\varepsilon}: L_2(\Gamma) \to V_h \subset L_2(\Omega)$. Then it is easy to verify that

$$(T_{h,\varepsilon}^* f_h, g) = (f_h, T_{h,\varepsilon}g), f_h \in V_h, g \in L_2(\Gamma)$$

is given by

$$T_{h,\varepsilon}^* f_h = \tilde{P}_h N_\varepsilon^* A_\varepsilon A_{h,\varepsilon}^{-1} f_h = \frac{1}{\varepsilon}\tilde{P}_h \beta A_{h,\varepsilon}^{-1} f_h$$

where in the last equality we used (2.8).

Now we recall (4.27) in Lemma 4.5 to claim that

$$\|T^*_{h,\varepsilon} f_h\|_{H^{1/2-\rho}(\Gamma)} \leqq C\|f_h\| \begin{cases} 1 & \text{if } V_h^0 \subset V_h, \\ \left(1+\dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right), & r \geqq 1+\dfrac{\sigma-1}{2\rho}, \end{cases}$$

and hence by the duality

$$\|T_{h,\varepsilon} g\| \leqq C|g|_{H^{-1/2+\rho}(\Gamma)} \begin{cases} 1 & \text{if } V_h^0 \subset V_h, \\ \left(1+\dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right), & r \geqq 1+\dfrac{\sigma-1}{2\rho} \end{cases}$$

which in view of (5.4) is exactly (5.3).

The proof of parts (ii) and (iii) of Theorem 3.1 is thus completed. For the proof of part (i) we use similar arguments. Indeed, first note that

$$D(A_\varepsilon^\alpha) = H^{2\alpha}(\Omega) \alpha < \tfrac{1}{4}. \tag{5.5}$$

In fact,

$$D(A^{1/2}) \subset D(A_\varepsilon^{1/2}).$$

Hence

$$H^{2\alpha}(\Omega) = D(A^\alpha) \subset D(A_\varepsilon^\alpha), \qquad \alpha < \tfrac{1}{4}. \tag{5.6}$$

On the other hand,

$$D(A_\varepsilon^{1/2}) \subset H^1(\Omega);$$

consequently,

$$D(A_\varepsilon^\alpha) \subset H^{2\alpha}(\Omega), \qquad \alpha < \tfrac{1}{2}. \tag{5.7}$$

Formula (5.7) combined with (5.6) yields (5.5).

By (5.5) and (4.45) in Lemma 4.6 applied with $\alpha < \frac{1}{4}$, we obtain

$$\|C_{h,\varepsilon}(t) v_h\|_{H^{2\alpha}(\Omega)} \leqq C\|v_h\|_{H^{2\alpha}(\Omega)}, \qquad \alpha < \tfrac{1}{4}.$$

Thus, to prove part (i) of Theorem 3.1 (recall (5.1)), it is enough to show that

$$\|A_{h,\varepsilon}^{-1} P_h A_\varepsilon N_\varepsilon \tilde{P}_h g\|_{H^{1/2}(\Omega)} \leqq C|g|_{H^\rho(\Gamma)}. \tag{5.8}$$

On the other hand, (5.8) follows easily from a dual version of (4.26) in Lemma 4.5 applied with $s = \frac{1}{2} - \rho$. The proof of Theorem 3.1 is thus completed. □

Next we prove Theorem 3.2.

Proof of Theorem 3.2. Let $\bar{u}_\varepsilon$ (respectively, $u_{h,\varepsilon}$) be the solution to (1.4) with $g \equiv \tilde{P}_h g$ (respectively, (1.5)). Then

$$((\ddot{\bar{u}}_\varepsilon - \ddot{u}_{h,\varepsilon})(t), \phi_h) + a_\varepsilon(\bar{u}_\varepsilon(t) - u_{h,\varepsilon}(t), \phi_h) = 0, \qquad \phi_h \in V_h.$$

Let $e_{h,\varepsilon}(t) \equiv P_{h,\varepsilon}\bar{u}_\varepsilon(t) - u_{h,\varepsilon}(t) \in V_h$. Then

$$\begin{aligned} &(\ddot{e}_{h,\varepsilon}(t), \phi_h) + a_\varepsilon(e_{h,\varepsilon}(t), \phi_h) = (P_{h,\varepsilon}\ddot{\bar{u}}_\varepsilon(t) - \ddot{\bar{u}}_\varepsilon(t), \phi_h), \\ &e_{h,\varepsilon}(0) = \dot{e}_{h,\varepsilon}(0) = 0. \end{aligned} \tag{5.9}$$

Equations (5.9) can be equivalently rewritten as

$$\ddot{e}_{h,\varepsilon}(t)+A_{h,\varepsilon}e_{h,\varepsilon}(t)=P_h(P_{h,\varepsilon}-I)\ddot{u}_\varepsilon(t),$$

$$e_{h,\varepsilon}(0)=\dot{e}_{h,\varepsilon}(0)=0,$$

or in the semigroup form as

$$e_{h,\varepsilon}(t)=\int_0^t S_{h,\varepsilon}(t-z)P_h(P_{h,\varepsilon}-I)\ddot{u}_\varepsilon(z)\,dz.$$

After integrating the above expression by parts and using $S_{h,\varepsilon}(0)=0$, we obtain

$$e_{h,\varepsilon}(t)=-S_{h,\varepsilon}(t)P_h(P_{h,\varepsilon}-I)\dot{u}_\varepsilon(0)-\int_0^t C_{h,\varepsilon}(t-z)P_h(P_{h,\varepsilon}-I)\dot{u}_\varepsilon(z)\,dz. \tag{5.10}$$

Since $\dot{u}_\varepsilon(0)=0$, (4.43) in Lemma 4.6 implies

$$\|e_{h,\varepsilon}(t)\|\leqq C\|P_h(P_{h,\varepsilon}-I)\dot{u}_\varepsilon\|_{L_1[0,T;L_2(\Omega)]}$$

(by the stability of orthogonal projection P_h)

$$\leqq C\|(P_{h,\varepsilon}-I)\dot{u}_\varepsilon\|_{L_1[0,T;L_2(\Omega)]}.$$

Now we will apply Lemma 4.2′. In fact, (4.23b) and (4.24b) in Lemma 4.2′ give

$$\|e_{h,\varepsilon}(t)\|\leqq Ch^{s+1}\left\{\begin{array}{l}1 \quad \text{if } V_h^0\subset V_h, \quad s+1\leqq r\\ h^{-2\rho}\left(1+\dfrac{h^{\sigma/2-1}}{2\rho}\right)^2, \quad r\geqq 1+\dfrac{s(\sigma-1)}{2\rho}, \quad 0\leqq s\leqq r-1\end{array}\right\} \tag{5.11}$$

$$\cdot\left[\|\dot{u}_\varepsilon\|_{L_1[0,T;H^{s+1}(\Omega)]}+|\dot{g}_h|_{L_1[0,T;H^{s+1}(\Gamma)]}+\sqrt{\varepsilon}\left|\frac{\partial}{\partial\eta}\dot{u}_\varepsilon\right|_{L_1[0,T;H^s(\Gamma)]}\right].$$

On the other hand, if we set

$$z_\varepsilon\equiv\dot{u}_\varepsilon(t),$$

then

$$\ddot{z}_\varepsilon(t)=A(x,\partial)z \quad \text{in } Q,$$

$$\left\{\begin{array}{l}z_\varepsilon(0)=\dot{u}_\varepsilon(0)=0\\ \dot{z}_\varepsilon(0)=A(x,\partial)\ddot{u}_\varepsilon(0)=0\end{array}\right. \quad \text{in } Q, \tag{5.12}$$

$$\frac{\varepsilon\partial z_\varepsilon}{\partial\eta}+\beta z_\varepsilon=\beta\dot{g} \quad \text{in } \Sigma.$$

Theorem 2.2(e) and (g) applied to (5.12) yield, in particular (provided $\dot{g}$ satisfies the appropriate compatibility conditions), that

$$\|z_\varepsilon\|_{L^\infty[0,T;H^{s+1}(\Omega)]}+\left|\frac{\partial z_\varepsilon}{\partial\eta}\right|_{L_2[0,T;H^s(\Gamma)]}\leqq C|\dot{g}_h|_{H^{s+1,s+1}(\Sigma)}\leqq C|\dot{g}|_{H^{s+1,s+1}(\Sigma)}. \tag{5.13}$$

Formula (5.13) together with (5.11), applied with $s\equiv s-1$, gives

$$\|e_{h,\varepsilon}(t)\|_{L_2(\Omega)}\leqq Ch^s|\dot{g}|_{H^{s,s}(\Sigma)}f(h,\rho,\sigma,\varepsilon) \tag{5.14}$$

where

$$f(h,\rho,\sigma,\varepsilon)\equiv\begin{cases}1 & \text{if } V_h^0\subset V_h,\quad 1\leqq s\leqq r\\ h^{-2\rho}\left[1+\dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right]^2, & r\geqq 1+\dfrac{(s-1)(\sigma-1)}{2\rho},\quad 1\leqq s.\end{cases}$$

Writing

$$u_\varepsilon-u_{h,\varepsilon}=u_\varepsilon-\bar{u}_\varepsilon+(I-P_{h,\varepsilon})\bar{u}_\varepsilon+e_{h,\varepsilon} \tag{5.15}$$

and applying part (a) of Theorem 2.2 together with the approximation properties for $\tilde{P}_h$ to the first term in (5.15), (4.23b) and (4.24b) to the second term, and (5.14) to the third term gives parts (i) and (iii) of Theorem 3.2. Finally, we will prove parts (ii) and (iv) of Theorem 3.2.

Recalling (5.10) and (4.45), and applying (5.5) with $\alpha=\frac{1}{2}-\rho$ yields

$$\|e_{h,\varepsilon}(t)\|_{H^{1/2-\rho}(\Omega)}\leqq C\|(P_{h,\varepsilon}-I)\dot{u}_\varepsilon\|_{L_1[0,T;H^{1/2-\rho}(\Omega)]}. \tag{5.16}$$

From (4.22b), (4.23c), and (4.24c) applied to the right-hand side of (5.16), we obtain

$$\begin{aligned}\|e_{h,\varepsilon}(t)\|_{H^{1/2-\rho}(\Omega)}&\leqq Ch^{1/2+s}\cdot k(h,\rho,\sigma,\varepsilon)\\ &\cdot\left[\|\dot{u}_\varepsilon\|_{L_1[0,T;H^{s+1}(\Omega)]}+|\dot{g}_h|_{L_1[0,T;H^{s+1}(\Gamma)]}+\sqrt{\varepsilon}\left|\frac{\partial\dot{u}_\varepsilon}{\partial\eta}\right|_{L_1[0,T;H^{s-1}(\Gamma)]}\right]\end{aligned} \tag{5.17a}$$

where

$$k(h,\rho,\sigma,\varepsilon)\equiv\begin{cases}h^{-\rho} & \text{if } s<\frac{1}{2}\\ 1 & \text{if } V_h^0\subset V_h,\quad s\leqq r-\frac{3}{2}\\ h^{-2\rho}\left(1+\dfrac{h^{\sigma/2-1}}{\sqrt{\varepsilon}}\right), & r\geqq 1+\dfrac{s(\sigma-1)}{2\rho}.\end{cases} \tag{5.17b}$$

Theorem 2.2(e)-(g) applied to (5.12) (recall $z_\varepsilon\equiv\dot{u}_\varepsilon$) gives

$$\|\dot{u}_\varepsilon\|_{L_1[0,T;H^{s+1}(\Omega)]}+\sqrt{\varepsilon}\left|\frac{\partial\dot{u}_\varepsilon}{\partial\eta}\right|_{L_1[0,T;H^{s-1}(\Gamma)]}\leqq C|\dot{g}_h|_{H^{s+1}(\Sigma)}\leqq C|\dot{g}|_{H^{s+1}(\Sigma)}. \tag{5.18}$$

Combining (5.17) and (5.18) yields

$$\|e_{h,\varepsilon}(t)\|_{H^{1/2-\rho}(\Omega)}\leqq Ch^{1/2+s}|\dot{g}|_{H^{s+1}(\Sigma)}\cdot k(h,\rho,\sigma,\varepsilon) \tag{5.19}$$

where $k(h,\rho,\sigma,\varepsilon)$ is given by (5.17b).

To complete the proof of parts (ii) and (iv) of Theorem 3.2, we apply the triangle inequality to (5.15). To estimate the first term on the right-hand side of (5.15) we interpolate between parts (a) and (b) of Theorem 2.2. This gives

$$\begin{aligned}\|u_\varepsilon-\bar{u}_\varepsilon\|_{C[0,T;H^{1/2-\rho}(\Omega)]}&\leqq\|L_\varepsilon(g-\tilde{P}_hg)\|_{C[0,T;H^{1/2-\rho}(\Omega)]}\\ &\leqq C|g-\tilde{P}_hg|_{H^{1/2-\rho}(\Sigma)}\\ &\leqq Ch^{1/2+s-\rho}|g|_{H^{s+1-2\rho}(\Sigma)}.\end{aligned} \tag{5.20}$$

To estimate the second term on the right-hand side of (5.15) we again use (4.22b), (4.23c), (4.24c) together with the regularity properties of $\bar{u}_\varepsilon$ as stated in Theorem 2.2. This gives

$$\|(I-P_{h,\varepsilon})\bar{u}_\varepsilon(t)\|_{H^{1/2-\rho}(\Omega)}\leqq Ch^{1/2+s}\cdot k(h,\rho,\sigma,\varepsilon)\cdot[|g|_{H^{s+1}(\Sigma)}+|\dot{g}|_{H^{s+1}(\Sigma)}]. \tag{5.21}$$

Collecting the results of (5.19), (5.20), (5.21) and (5.15) completes the proof of parts (ii) and (iv) of Theorem 3.2. □

Acknowledgment. The authors thank Professor I. Babuška for pointing out and making available a preprint of [B4].

REFERENCES

[B1] I. BABUŠKA AND A. AZIZ, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972.

[B2] I. BABUŠKA, *The finite element method with penalty*, Math. Comp., 27 (1973), pp. 221-228.

[B3] ———, *Approximation by Hill functions* II, Comment. Math. Univ. Carolin., 13 (1972), pp. 1-22.

[B4] I. BABUŠKA AND M. SURI, *The h-p version of the finite element method with quasiuniform meshes*, Tech. Report BN-1046, Institute for Physical Science and Technology, University of Maryland, College Park, MD, 1986.

[B5] G. BAKER, *Error estimates for finite element methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 13 (1976), pp. 564-576.

[B6] L. BALES, *Higher order single step fully discrete approximations for second order hyperbolic equations with time dependent coefficients*, SIAM J. Numer. Anal., 23 (1986), pp. 27-43.

[B7] J. J. BLAIR, *Higher order approximations to the boundary conditions for the finite element method*, Math. Comp., 30 (1976), pp. 250-262.

[D1] J. DOUGLAS AND T. DUPONT, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 7 (1970), pp. 575-626.

[D2] T. DUPONT, *L^2-estimates for Solerkin methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 880-889.

[F1] H. FATTORINI, *Second Order Linear Differential Equations in Banach Spaces*, North-Holland, Amsterdam, 1985.

[L1] J. L. LIONS, *A remark on the approximation of nonhomogenous hyperbolic boundary value problems*, paper dedicated to G. I. Marchuk on his sixtieth anniversary, Vistas in Applied Mathematics, A. V. Balakrishnan, A. A. Dorodnicen, J. L. Lions, eds., Optimization Software, New York, 1986.

[L2] — —, *Contrôle des systems distribués singuliers*, Dunod, Paris, 1983.

[LM] J. L. LIONS AND E. MAGENES, *Nonhomogenous Boundary Value Problems and Applications*, Vols. I and II, Springer-Verlag, Berlin, New York, 1972.

[LLT] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Nonhomogenous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986), pp. 149-192.

[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.

[LS1] I. LASIECKA AND J. SOKOLOWSKI, *Regularity and strong convergence of variational approximation to a nonhomogenous Dirichlet hyperbolic boundary value problem*, SIAM J. Math. Anal., 19 (1988), pp. 528-540.

[LS2] I. LASIECKA, J. SOKOLOWSKI, AND P. NEITTAANMAKI, *Finite element approximations of the wave equation with Dirichlet boundary data on a bounded domain in R^2*, in Proc. Conference on Control of Distributed Parameter Systems, Voran, Austria, July 1986.

[P1] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.

[R1] J. RAUCH, *On convergence of the finite element method for the wave equation*, SIAM J. Numer. Anal., 22 (1985), pp. 245-250.

[N1] J. A. NITSCHE, *Uber ein Variationsprinzip zur Lösung von Dirichlet-problemen bei Verwendung von Teilraumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9-15.

[S1] R. SCOTT, *Interpolated boundary conditions in the finite element method*, SIAM J. Numer. Anal., 12 (1975), pp. 404-427.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1388-1425, November 1989

007

UNIQUENESS OF SOLUTIONS FOR THE GENERALIZED KORTEWEG-DE VRIES EQUATION*

J. GINIBRE† AND Y. TSUTSUMI‡

Abstract. The uniqueness of L^2 and H^1 solutions of the Cauchy problem for the generalized Korteweg-de Vries (KdV) equation

$$\partial_t u + D^3 u + a(u)Du = 0$$

with $a \in \mathscr{C}(\mathbb{R}, \mathbb{R})$ and with initial data u_0 in weighted L^2 or H^1 spaces according to

$$(1+x_+)^{\beta/2} u_0 \in L^2, \qquad (1+x_+)^{\gamma/2} Du_0 \in L^2$$

is studied for some $\beta, \gamma \geqq 0$. Several uniqueness classes are exhibited, and the a priori estimates are derived for the corresponding norms of smooth solutions in terms of the initial data required to implement compactness existence proofs of solutions in those classes. For (weighted) L^2 solutions, the results given here cover the case where $|a(\rho)| \leqq C|\rho|^p$ with $0 < p < \frac{7}{2}$, $\beta = 1/p - \frac{1}{4}$ if $p \leqq 2$ and $\beta = \frac{1}{4}$ if $p \geqq 2$. For the ordinary KdV equation with $p = 1$, the result $\beta = \frac{3}{4}$ improves over previously known results by a factor of 2. For H^1 solutions, uniqueness and a priori estimates with initial data $u_0 \in H^1$ (namely, $\beta = \gamma = 0$) are proved provided $p > \frac{3}{2}$. For the ordinary KdV equation with $p = 1$, the results given here yield uniqueness and a priori estimates for $\beta + \gamma \geqq \frac{1}{2}$, $\beta \geqq 3\gamma/5$ (for instance, $\beta = \frac{1}{2}$, $\gamma = 0$, or $\beta = (3/16)$, $\gamma = (5/16)$).

Key words. Korteweg-de Vries (KdV) equation, Cauchy problem, uniqueness of solutions

AMS(MOS) subject classifications. primary 35Q20; secondary 35G25, 35D99

1. Introduction.

In this paper we study the uniqueness of solutions of the Cauchy problem for the generalized Korteweg-de Vries (GKdV) equation

$$\partial_t u + D^3 u = DV'(u) \tag{1.1}$$

for $t \geqq 0$, $x \in \mathbb{R}$, with initial data

$$u(0, x) = u_0(x) \qquad (x \in \mathbb{R}). \tag{1.2}$$

Here $D = d/dx$, the prime denotes the derivative, and $V \in \mathscr{C}^2(\mathbb{R}, \mathbb{R})$ with $V(0) = V'(0) = 0$. The function V is the potential that appears in the (formally conserved) energy for (1.1) (see (4.1) below). In the notation of [10], (1.1) is written equivalently as

$$\partial_t u + D^3 u + a(u)Du = 0 \tag{1.1'}$$

with $a = -V''$. For $a(\rho) = \rho$, (1.1') is the (ordinary) Korteweg-de Vries equation, and for $a(\rho) = \rho^2$, (1.1') is the modified Korteweg-de Vries equation.

A large amount of work has been devoted to the existence problem for solutions of (1.1), (1.2). It concerns either the ordinary KdV equation with $a(\rho) = \rho$ [2]-[5], [12], [16]-[18] or the generalized equation (1.1') with a monomial or a smooth function [1], [9], [10], [15], [19]-[21]. The available results include, in particular, the existence of global weak solutions in $L^\infty(\mathbb{R}, L^2)$ for initial data $u_0 \in L^2$ and in $L^\infty(\mathbb{R}, H^1)$ for initial data $u_0 \in H^1$. They require suitable growth restrictions on the function a at

* Received by the editors July 1, 1988; accepted for publication November 18, 1988. This work was written while the second author was visiting the Laboratoire de Physique Théorique et Hautes Energies, which is affiliated with the Centre National de la Recherche Scientifique.

† Laboratoire de Physique Théorique et Hautes Energies, Université de Parix XI, Bâtiment 211, 91405 Orsay, Cedex, France.

‡ Faculty of Integrated Arts and Science, Hiroshima University, Higashisenda-machi, Naka-ku, Hiroshima 730, Japan.

infinity. There also exists a wealth of results concerning more regular solutions, for instance, solutions that are continuous functions of time with values in Sobolev spaces H^s for $s>\frac{3}{2}$ corresponding to initial data in H^s. We refer to [10] and [14] for a general survey.

The problem of the uniqueness of the solutions of (1.1), (1.2), on the other hand, is still largely open. There exists a well-known result according to which the solution is unique if $u_0\in H^s$ with $s>\frac{3}{2}$ (see [2], [9], [16]), but that result does not cover the case of H^1 solutions. More recently, new uniqueness results were proved in the special case $a(\rho)=\rho$ of the ordinary KdV equation for initial data u_0 in weighted L^2 spaces with either exponential [10] or polynomial [12] weight. The method of proof uses the smoothing properties of the KdV equation, and propagation in time of space decay. It is not, however, directly applicable to cases other than $a(\rho)=\rho$.

In the present paper, we study the uniqueness problem for L^2 or H^1 solutions of (1.1), (1.2)—more precisely, solutions in $L^\infty_{\text{loc}}([0, T), L^2)$ or in $L^\infty_{\text{loc}}([0, T), H^1)$ for some $T>0$ (possibly $T=\infty$), for initial data u_0 in (possibly weighted) L^2 or H^1 spaces, more precisely, for u_0 satisfying

$$(1+x_+)^{\beta/2}u_0\in L^2, \tag{1.3}$$

$$(1+x_+)^{\gamma/2}Du_0\in L^2, \tag{1.4}$$

for some $\beta, \gamma\geqq 0$, where $x_+=\max(x, 0)$. In particular, we obtain several uniqueness classes for solutions of (1.1), (1.2) under suitable assumptions of β, γ, and V (or a). Of course, uniqueness results of this kind are of interest only insofar as the uniqueness classes thereby obtained are suitable for proving the existence of solutions. One possible method for proving existence consists of first constructing smooth solutions of a regularized version of problem (1.1), (1.2), then deriving a priori estimates of the solutions in terms of the initial data, and finally exploiting those estimates to remove the regularization by a limiting procedure based on compactness arguments. The limiting procedure, in general, takes a weak-star limit in a space X^* that is the dual of some Banach space X; the a priori estimates, which hold for the X^* norm of the regularized solutions, carry over to the limit, so that X^* turns out to be an adequate space for the existence of solutions.

In the present paper we do not consider the problem of existence. However, we provide the technical material required to implement the previous type of existence proof, by deriving a priori estimates of sufficiently smooth solutions of (1.1), (1.2), in the norms that define the previous uniqueness classes, in terms of the available norms of the initial data. Those estimates strongly indicate that the previous uniqueness classes are in fact suitable for existence and uniqueness insofar as they are duals of Banach spaces. That indication can be made into a proof with little additional work by combining our results with many of the available existence results. However, we refrain from making any formal statement to that effect in this paper, because the smoothness assumptions that we make on V (or a) are weaker than the assumptions made in the existence proofs available in the literature. In a subsequent paper, we will prove the existence of solutions of (1.1) under weaker assumptions on V than those of the present paper, and extend the uniqueness proofs of this paper into existence and uniqueness proofs in the uniqueness classes obtained here.

We now give a rough outline of our results, namely, uniqueness and a priori estimates, as just explained, concentrating on the assumptions on V (or a) and on u_0. More precise statements, in particular, the description of the uniqueness classes, will be found in the main body of the paper, whose contents are described below.

In the case of L^2 solutions, the assumption on V (or a) reduces to

$$|a(\rho)| \leqq C|\rho|^p \tag{1.5}$$

for all $\rho \in \mathbb{R}$ and some $p > 0$, and the assumption on u_0 reduces to (1.3). Our results cover the case $0 < p \leqq 2$ with $\beta = 1/p - \frac{1}{4}$, and, by the use of two different methods, either $2 \leqq p < 3$ with $\beta = \frac{1}{2} - 1/(2p)$ with the simpler one, or $2 \leqq p < \frac{7}{2}$ with $\beta = \frac{1}{4}$ with the more elaborate one. In particular, for the ordinary KdV equation with $p = 1$, we obtain $\beta = \frac{3}{4}$, thereby improving the corresponding result of [12] by a factor of 2. For the modified KdV equation, we obtain the lower value $\beta = \frac{1}{4}$.

In the case of H^1 solutions, the assumptions on V (or a) reduce to condition (1.5), but now only for small ρ (say for $|\rho| \leqq 1$), to the condition that a be absolutely continuous with locally bounded derivative, and to a lower semiboundedness condition on V (see (4.22) below). The assumptions on u_0 include (1.3) and (1.4). Our results cover the cases where $p \geqq 1$, $\beta \geqq \gamma(4-p)_+/(4+p)$ with either of the two conditions

$$(p+1)(\beta+\gamma) \geqq 1, \tag{1.6}$$

$$(3-p)_+\beta + (\rho-1)(2+3\min(\beta, \gamma)) > 1. \tag{1.7}$$

In particular by (1.7) we prove uniqueness for $u_0 \in H^1$ provided $p > \frac{3}{2}$, thereby recovering a result previously obtained by Tsutsumi [22]. That result covers, in particular, the case of the modified KdV equation corresponding to $p = 2$. For the ordinary KdV equation with $p = 1$, we cannot prove uniqueness for $u_0 \in H^1$ only, but we obtain uniqueness (for instance) for $\beta = \frac{1}{2}$, $\gamma = 0$ or for $\beta = 3/(16)$, $\gamma = 5/(16)$, by using (1.6).

The proofs of our results rely on two sets of estimates. The first set consists of the well-known weighted L^2, H^1, and H^2 estimates that have been widely used by many previous authors. Those estimates can in part be viewed as weighted space time integrability properties associated with the free evolution group $U(t) = \exp(-tD^3)$ that solves the linear equation obtained from (1.1) when $V = 0$. By interpolation, they can be combined with similar weighted L^∞ estimates coming from known estimates of the propagator of $U(t)$, which is the classical Airy function. The second set of estimates consists of space time integrability properties that stand in close analogy with similar properties of other equations, and in particular of the Schrödinger equation [7], [11], [23].

This paper is organized as follows. In § 2, we derive the linear estimates of $U(t)$ just described. The Schrödinger-like estimates are contained in Lemmas 2.1 and 2.2, the weighted L^∞ estimates in Lemmas 2.3 and 2.4, the weighted L^2 and H^1 estimates in Lemma 2.5, and the interpolation between the last two estimates in Lemma 2.6.

In § 3 we derive our results on L^2 solutions. We first obtain the basic uniqueness classes that are suitable for $p \leqq 2$ and $p \geqq 2$ in Propositions 3.1 and 3.2, respectively. We then set out to derive a priori estimates for the norms that appear in their definition. In both cases, one of those norms plays only an auxiliary role and can be estimated in terms of the others (Propositions 3.3 and 3.4, respectively). To proceed further, we use a well-known weighted L^2 identity (I.1) to derive weighted L^2 and other estimates (Proposition 3.5), following previous authors (see especially [10] and [12]). That step completes the proof of a priori estimates for $p \leqq 2$. For $p \geqq 2$, more work is needed and we offer two methods. One of them is based on Sobolev inequalities and covers the case $2 \leqq p < 3$ with $\beta = \frac{1}{2} - 1/(2p)$, yielding as a byproduct a new uniqueness class (Proposition 3.6). The second (more elaborate) method covers the case $2 \leqq p < \frac{7}{2}$ with $\beta = \frac{1}{4}$, and also yields a new uniqueness class (Proposition 3.7). We then address the problem of deriving the basic identity (I.1) under minimal smoothness assumptions and analyze that condition in some detail (Proposition 3.8). When combined with

Proposition 3.5, that study makes it possible to exhibit new uniqueness classes both for $p \leqq 2$ (Proposition 3.9) and for $p \geqq 2$ (Proposition 3.10), but those classes are probably not suitable for a treatment of the existence problem (see Remark 3.8).

In § 4, we derive our results on H^1 solutions. We first remark that weighted L^2 estimates can be obtained from (I.1) more simply and in stronger form than for L^2 solutions (Proposition 4.1). We then obtain the basic uniqueness class in Proposition 4.2, which is a mild generalization of a known result. To proceed further, we need a second, well-known, weighted H^1 identity (I.2). As it turns out, it is easier to derive that identity under natural smoothness assumptions than to derive (I.1) for L^2 solutions, and we analyze it in some detail (Proposition 4.3). Following previous authors, we derive therefrom weighted H^1 and other estimates (Proposition 4.4). Using those preliminaries, we turn to the derivation of a priori estimates of the norms that define the uniqueness class of Proposition 4.2. That can be done again by two methods. The first method leads to condition (1.6) (under assumptions (1.3), (1.4)) and yields as a byproduct a new uniqueness class (Proposition 4.5). The second (more elaborate) method leads to condition (1.7) and also yields a new uniqueness class (Proposition 4.6), slightly smaller than the previous one. Combining those results with Propositions 4.3 and 4.4 makes it possible to exhibit still another uniqueness class (Proposition 4.7), which, in contrast with the previous ones, is probably not suitable for a treatment of the existence problem if $p < 2$. We conclude that section with some comments on the assumptions of Propositions 4.5 and 4.6.

We conclude this Introduction by giving some notation that will be used without further explanation throughout this paper. We denote by $\|\cdot\|_r$ the norm in $L^r \equiv L^r(\mathbb{R})$, and by $\langle .,. \rangle$ the scalar product in L^2. Pairs of conjugate indices are written as r, $\bar{r}$, with $1/r + 1/\bar{r} = 1$. For any interval $I \subset \mathbb{R}$, for any Banach space X, we denote by $\mathscr{C}_w(I, X)$ the space of weakly continuous functions from I to X, and by $L^q(I, X)$ (respectively, $L^q_{\mathrm{loc}}(I, X)$) the space of measurable functions v from I to X such that $\|v(\cdot); X\| \in L^q(I)$ (respectively, $L^q_{\mathrm{loc}}(I)$). We will make extensive use of the following spaces. Let $1 \leqq q, r, s \leqq \infty$. For any $j \in \mathbb{Z}$, let χ_j be the characteristic function of the interval $[j - \frac{1}{2}, j + \frac{1}{2}]$. We define $l^s(L^q(I, L^r))$ as the space of functions v of space time defined for $(x, t) \in \mathbb{R} \times I$ for which the following quantity (taken as the norm) is finite:

$$\|v; l^s(L^q(I, L^r))\| = \left\{ \sum_{j \in \mathbb{Z}} \left[\int_I dt \left(\int dx \, |\chi_j v|^r \right)^{q/r} \right]^{s/q} \right\}^{1/s} < \infty.$$

We also define the local spaces $l^s(L^q_{\mathrm{loc}}(I, L^r))$ as the spaces of functions v defined in $\mathbb{R} \times I$ and such that for any compact interval $J \Subset I$, the restriction $v|_J$ belongs to $l^s(L^q(J, L^r))$. Similarly, we define the spaces $L^q(I, l^s(L^r))$ with norm

$$\|v; L^q(I, l^s(L^r))\| = \left\{ \int_I dt \left[\left(\sum_{j \in \mathbb{Z}} \int dx \, |\chi_j v|^r \right)^{s/r} \right]^{q/s} \right\}^{1/q} < \infty$$

and the corresponding local spaces $L^q_{\mathrm{loc}}(I, l^s(L^r))$. Clearly $L^q(I, l^s(L^r)) \equiv L^q(I, L^r)$ if $s = r$. By the Minkowski inequality the continuous embedding $l^s(L^q(I, L^r)) \subset L^q(I, l^s(L^r))$ holds if $s \leqq q$ and the converse embedding holds if $s \geqq q$. The Hölder and Young inequalities hold in all those spaces, as a consequence of the corresponding inequalities applied independently in each of the three component spaces.

In all the estimates performed in this paper, we use C to denote various constants, possibly depending on various indices $(\alpha, \beta, \gamma, p, q, r, s)$, but not on the functions to be estimated, and possibly different from one line to the next. In the entire paper, we assume without further repetition that the function V in (1.1) satisfies $V \in \mathscr{C}^2(\mathbb{R}, \mathbb{R})$, $V(0) = V'(0) = 0$. Additional assumptions, if any, will be stated where needed. Finally,

we denote by χ_+ and χ_- the characteristic functions of $\mathbb{R}^+$ and $\mathbb{R}^-$, respectively, and for any $\lambda \in \mathbb{R}$, we define $\lambda_\pm = \max(\pm\lambda, 0)$.

2. Linear estimates. In this section, we derive a number of preliminary estimates on the free evolution group

$$U(t) = \exp(-tD^3) \tag{2.1}$$

that will be essential in the rest of this paper. The group $U(\cdot)$ solves the Cauchy problem for the linear equation

$$\partial_t u + D^3 u = f, \tag{2.2}$$

with initial condition $u(0) = u_0$ through the formula

$$u(t) = U(t)u_0 + \int_0^t d\tau U(t-\tau) f(\tau). \tag{2.3}$$

Since all the estimates considered here are linear, it suffices to derive them for smooth functions. They can then be extended by continuity to all functions for which they make sense.

The group $U(\cdot)$ is unitary in L^2. For each t, $U(t)$ can be represented as the convolution with the function

$$S_t(x) = (3t)^{-1/3} \operatorname{Ai}(x(3t)^{-1/3}) \tag{2.4}$$

where Ai is the classical Airy function

$$\operatorname{Ai}(x) = (2\pi)^{-1} \int d\xi \exp(i\xi^3/3 + ix\xi). \tag{2.5}$$

It satisfies the estimates [8, p. 213]

$$|\operatorname{Ai}(x)| \leqq C(1+x_-)^{-1/4} \exp(-cx_+^{3/2}), \tag{2.6}$$

$$|\operatorname{Ai}'(x)| \leqq C(1+x_-)^{1/4} \exp(-cx_+^{3/2}). \tag{2.7}$$

From (2.4) and (2.6) we obtain an estimate

$$|S_t(x)| \leqq t^{-1/3} A(x) + Ct^{-1/4} \chi_-(1+|x|)^{-1/4} \tag{2.8}$$

for some smooth bounded function A with support in $[-1, \infty)$ and (faster than) exponential decrease at $+\infty$, uniformly for t in bounded intervals. Furthermore,

$$\|S_t\|_\infty = Ct^{-1/3}. \tag{2.9}$$

Similarly,

$$|DS_t(x)| \leqq t^{-2/3} A(x) + Ct^{-3/4} \chi_-(1+|x|)^{1/4}. \tag{2.10}$$

From the unitarity of U and from (2.9), we can derive the following properties.

LEMMA 2.1. *The following estimates hold*:

$$\|U(t)f\|_r \leqq Ct^{-(1/3-2/(3r))} \|f\|_{\bar{r}} \tag{2.11}$$

for all $t \neq 0$ *and all* r, $2 \leqq r \leqq \infty$,

$$\|U(\cdot)f\colon L^q(\mathbb{R}, L^r)\| \leqq C\|f\|_2 \tag{2.12}$$

for all r, $2 \leqq r \leqq \infty$, *with* $2/q = \frac{1}{3} - 2/(3r)$,

$$\left\| \int_I d\tau\, U(\cdot - \tau) f(\tau);\ L^{q_1}(\mathbb{R}, L^{r_1}) \right\| \leqq C\|f;\ L^{\tilde{q}_2}(I, L^{\tilde{r}_2})\| \tag{2.13}$$

for all r_i, $2 \leqq r_i \leqq \infty$, $i = 1, 2$, *with* $2/q_i = \frac{1}{3} - 2/(3r_i)$ *and for any interval* I.

Summary proof. Estimates (2.11)-(2.13) are the analogues of well-known estimates for the Schrödinger equation, and are proved in exactly the same manner. The estimate (2.11) follows by interpolation from the cases $r=2$ (unitarity of U) and $r=\infty$, where it follows from (2.9) through the Young inequality. We refer to [7, Prop. 4.4] for (2.12) and to [11] and [23] for (2.13). □

For later reference, we state explicitly the special cases $r=\infty$, $q=6$ of (2.12) and $r_1=\infty$, $q_1=6$, $r_2=2$, $q_2=\infty$ of (2.13):

$$\|U(\cdot)f;\, L^6(\mathbb{R}, L^\infty)\| \leqq C\|f\|_2, \tag{2.14}$$

$$\left\| \int_I d\tau\, U(\cdot-\tau)f(\tau);\, L^6(\mathbb{R}, L^\infty) \right\| \leqq C\|f;\, L^1(I, L^2)\|. \tag{2.15}$$

We also remark that the integration interval on the left-hand side of (2.13), (2.15) may also depend on the time variable. In that case, the interval on the right-hand side must be replaced by the union of the intervals that occur in the left-hand side.

We now turn to a similar set of estimates satisfied by the operator

$$e^x U(t) e^{-x} = \exp[-t(D-1)^3]. \tag{2.16}$$

LEMMA 2.2. *Let $T>0$. Then the following estimates hold:*

$$\|\exp[-t(D-1)^3]f\|_r \leqq C\, e^t t^{-((1/3)-(2/(3r)))}\|f\|_{\bar r} \tag{2.17}$$

for all $t>0$ and all r, $2\leqq r\leqq\infty$ (with the same C as in (2.11)),

$$\|\exp[-\cdot(D-1)^3]f;\, L^q([0, T], L^r]\| \leqq C\, e^T\|f\|_2 \tag{2.18}$$

for all r, $2\leqq r\leqq\infty$, with $2/q=\frac{1}{3}-2/(3r)$ (with the same C as in (2.12)),

$$\left\| \int_0^\cdot d\tau \exp[-(\cdot-\tau)(D-1)^3]f;\, L^{q_1}([0, T], L^{r_1}) \right\| \leqq C\, e^T\|f;\, L^{\bar q_2}([0, T], L^{\bar r_2})\| \tag{2.19}$$

for all r_i, $2\leqq r_i\leqq\infty$, $i=1, 2$, with $2/q_i=\frac{1}{3}-2/(3r_i)$ (with the same C as in (2.13)),

$$\left\| \int_0^\cdot d\tau \exp[-(\cdot-\tau)(D-1)^3]Df;\, L^q([0, T], L^\infty) \right\| \leqq C_q\, e^T T^{1/q-1/6}\|f;\, L^2([0, T], L^2)\| \tag{2.20}$$

for all q, $2\leqq q<6$ (C_q blows up as q increases to 6).

Proof. Estimates (2.17)-(2.19) follow immediately from estimates (2.11)-(2.13) if we expand $(D-1)^3$ in the exponential and note that the operator $\exp(-3tD)$ is the operator of translation by $3t$ and is therefore isometric in L^r for all r, while the diffusion operator $\exp(3tD^2)$ is a contraction in L^r for all r, $1\leqq r\leqq\infty$, and all $t\geqq 0$.

We now turn to the proof of (2.20). To prove that the operator K defined by

$$Kf(t) = \int_0^t d\tau \exp[-(t-\tau)(D-1)^3]Df(\tau) \tag{2.21}$$

is bounded from $L^2([0, T], L^2)$ to $L^q([0, T], L^\infty)$, it suffices to prove that the adjoint operator in $L^2([0, T], L^2)$, namely, the operator K^* defined by

$$K^*f(t) = -\int_t^T d\tau \exp[(\tau-t)(D+1)^3]Df(\tau) \tag{2.22}$$

is bounded from $L^{\bar{q}}([0, T], L^1)$ to $L^2([0, T], L^2)$. For that purpose we compute

$$\begin{aligned}&\|K*f; L^2([0, T], L^2)\|^2\\ &= \int_0^T dt \int_t^T d\tau\, d\tau' \langle Df(\tau), \exp[(\tau'-\tau)(D^3+3D)\\ &\qquad +(\tau+\tau'-2t)(3D^2+1)]Df(\tau')\rangle.\end{aligned} \tag{2.23}$$

Now the operator $D\exp(3tD^2)$ is represented by the convolution with the function

$$(-x/6t)(12\pi t)^{-1/2}\exp(-x^2/(12t))$$

and is therefore bounded in L^r for all r, $1\leqq r\leqq\infty$, by the Young inequality, with

$$\|\exp(3tD^2)Df\|_r \leqq (3\pi t)^{-1/2}\|f\|_r. \tag{2.24}$$

Using (2.11) with $r=\infty$ and (2.24), we estimate

$$\begin{aligned}&\|K*f; L^2([0, T], L^2)\|^2\\ &\leqq C\int_0^T dt \int_t^T d\tau\, d\tau' |\tau-\tau'|^{-1/3}\left(\frac{\tau+\tau'}{2-t}\right)^{-1}\exp(\tau+\tau'-2t)\|f(\tau)\|_1\|f(\tau')\|_1\\ &\leqq Ce^{2T}\int_0^T d\tau\, d\tau' |\tau-\tau'|^{-1/3}\|f(\tau)\|_1\|f(\tau')\|_1\int_0^{\min(\tau,\tau')} dt\left(\frac{\tau+\tau'}{2-t}\right)^{-1}.\end{aligned} \tag{2.25}$$

We perform the last integral and continue (2.25) as

$$\begin{aligned}&\cdots\leqq Ce^{2T}\int_0^T d\tau\, d\tau' |\tau-\tau'|^{-1/3}\log(2T|\tau-\tau'|^{-1})\|f(\tau)\|_1\|f(\tau')\|_1\\ &\leqq Ce^{2T}T^{2/q-1/3}\||\cdot|^{-1/3}\log(2|\cdot|^{-1}); L^{q/2}([-1, 1])\|f; L^{\bar{q}}([0, T], L^1)\|^2\end{aligned} \tag{2.26}$$

for all q, $2\leqq q<6$, by the Young inequality and an elementary homogeneity argument. The estimate (2.20) follows from (2.26) by duality, as explained above, and the constant C_q can be read off from (2.26). □

We now turn to the derivation of a set of weighted estimates for the operator $U(\cdot)$. The first estimates of this type are in L^∞ and rely on a direct use of (2.7).

LEMMA 2.3. *Let h be a continuous nonnegative function satisfying*

$$\begin{aligned}&h(x)=\exp(\alpha x) \quad \textit{for some } \alpha>0 \textit{ and all } x\leqq 0,\\ &0\leqq h(y)\leqq h(x)\leqq\exp[\alpha(x-y)]h(y) \quad \textit{for all } x\geqq y.\end{aligned} \tag{2.27}$$

Then the following estimate holds:

$$h^{1/2}(x)|DS_t(x-y)|h^{-1/2}(y)\leqq\begin{cases} C(t)t^{-3/4}(1+y)^{1/4} & \textit{for } y\geqq x_+,\\ C(t)t^{-3/4}\exp(-\gamma|x-y|/2) & \textit{for } y\leqq x_+\end{cases} \tag{2.28}$$

for $0\leqq\gamma<\alpha$, where $C(t)$ is uniformly bounded in t on bounded intervals and in γ for $0\leqq\gamma\leqq\gamma_0<\alpha$.

Proof. (C and c in this proof denote absolute constants, possibly different from one line to the next.) From (2.4), (2.7) we obtain

$$|DS_t(x-y)|\leqq Ct^{-3/4}(t^{1/12}+(y-x)_+^{1/4})\exp[-ct^{-1/2}(x-y)_+^{3/2}]. \tag{2.29}$$

We consider different cases. For $y\geqq x\geqq 0$, we use the fact that $h(x)\leqq h(y)$ to obtain

$$h^{1/2}(x)|DS_t(x-y)|h^{-1/2}(y)\leqq Ct^{-3/4}(t^{1/2}+y^{1/4}). \tag{2.30}$$

For $y \geqq 0 \geqq x$, we use the fact that $h(y) \geqq 1$, $h(x) = \exp(\alpha x)$ to obtain

$$h^{1/2}(x)|DS_t(x-y)|h^{-1/2}(y) \leqq Ct^{-3/4}(t^{1/12} + y^{1/4} + \sup_{x \leqq 0} |x|^{1/4} e^{\alpha x/2}) \tag{2.31}$$
$$= Ct^{-3/4}(t^{1/12} + y^{1/4} + (2\alpha e)^{-1/4})$$

by an elementary computation. For $y \leqq x$ or $x \leqq y \leqq 0$, we use the fact that $h(x)h^{-1}(y) \leqq \exp[\alpha(x-y)]$. Combining that estimate with (2.29), we obtain for $y \leqq x$

$$h^{1/2}(x)|DS_t(x-y)|h^{-1/2}(y) \leqq Ct^{-2/3} \exp(-\gamma|x-y|/2)$$
$$\cdot \sup_{x \geqq y} \exp[(\alpha+\gamma)(x-y)/2 - ct^{-1/2}(x-y)^{3/2}] \tag{2.32}$$
$$= Ct^{-2/3} \exp[-\gamma|x-y|/2 + ct(\alpha+\gamma)^3].$$

Finally, for $x \leqq y \leqq 0$ and $\gamma < \alpha$,

$$h^{1/2}(x)|DS_t(x-y)|h^{-1/2}(y) \leqq Ct^{-3/4} \exp(\gamma(x-y)/2)$$
$$\cdot \left\{ t^{1/12} + \sup_{x \leqq y} (y-x)^{1/4} \exp[(\alpha-\gamma)(x-y)/2] \right\} \tag{2.33}$$
$$= Ct^{-3/4} \exp(\gamma(x-y)/2)[t^{1/12} + (2(\alpha-\gamma)e)^{-1/4}].$$

Collecting (2.30)-(2.33) yields (2.28) with

$$C(t) = C\{1 + t^{1/12} \exp[ct(\alpha+\gamma)^3] + (\alpha-\gamma)^{-1/4}\}. \tag{2.34}$$

This proves the lemma. □

As an immediate consequence of Lemma 2.3, we obtain the following weighted $L^1 - L^\infty$ estimates for the operator U.

LEMMA 2.4. *Let h be a continuous function satisfying* (2.27). *Then the following estimates hold*:

$$\|h^{1/2}U(t)Dv\|_\infty \leqq C(t)t^{-3/4}\|h^{1/2}(1+x_+)^{1/4}v\|_1, \tag{2.35}$$

$$\left\| h^{1/2} \int_0^\cdot d\tau\, U(\cdot - \tau)Df(\tau);\ L^q([0,T], L^\infty) \right\| \tag{2.36}$$
$$\leqq C(T)\|h^{1/2}(1+x_+)^{1/4}f;\ L^l([0,T], L^1)\|,$$

for all q, l with $0 < 1/q = 1/l - \frac{1}{4} < \frac{3}{4}$ or for $q = \infty$, $l > 4$, and with $C(\cdot)$ uniformly bounded on the compact subsets of $\mathbb{R}^+$.

Proof. Inequalities (2.35) and (2.36) follow from (2.28) with $\gamma = 0$. Inequality (2.36) requires in addition the use of the Hardy-Littlewood-Sobolev inequality [8, p. 117] or of the Hölder inequality in time, depending on whether $q < \infty$ or $q = \infty$. □

We now turn to the derivation of weighted L^2-estimates. The basic result is the following.

LEMMA 2.5. *Let either $h \in \mathscr{C}^3(\mathbb{R}, \mathbb{R}^+)$ satisfy $h' \geqq 0$ and $h''' \leqq ch$ or $h \in \mathscr{C}(\mathbb{R}, \mathbb{R}^+)$ satisfy $h' \geqq 0$ and $h''^2 \leqq c^2 hh'$. For all f such that $h^{1/2}f \in L^1_{\mathrm{loc}}([0,T), L^2)$ and all t, $0 \leqq t < T$, define*

$$g(t) = \int_0^t d\tau \exp[c(t-\tau)/2]\|h^{1/2}f(\tau)\|_2. \tag{2.37}$$

Then the following estimates hold for all $t \geqq 0$, $t > 0$:

$$\|h^{1/2}U(t)u_0\|_2 \leqq e^{ct/2}\|h^{1/2}u_0\|_2, \tag{2.38}$$

$$\left\| h^{1/2} \int_0^t d\tau\, U(t-\tau)f(\tau) \right\|_2 \leqq g(t), \tag{2.39}$$

$$\left\| h^{1/2} \int_0^{\cdot} d\tau\, U(\cdot - \tau) f(\tau);\, L^\infty([0, T], L^2) \right\| \leqq e^{cT/2} \| h^{1/2} f;\, L^1([0, T], L^2) \|, \tag{2.40}$$

$$\| h'^{1/2} U(\cdot) D u_0;\, L^2([0, T], L^2) \| \leqq 2^{-1/2}\, e^{cT/2} \| h^{1/2} u_0 \|_2, \tag{2.41}$$

$$\left\| h'^{1/2} \int_0^{\cdot} d\tau\, U(\cdot - \tau) D f(\tau);\, L^2([0, T], L^2) \right\| \leqq 2^{-1/2} g(T) \leqq 2^{-1/2}\, e^{cT/2} \| h^{1/2} f;\, L^1([0, T], L^2) \|. \tag{2.42}$$

Proof. Let u be defined by (2.3) for smooth u_0 and f. From the differential equation (2.2) and the commutation relation

$$[D^3, h] = 3Dh'D + [D, h''] = 3Dh'D + h''', \tag{2.43}$$

possibly followed by the Schwarz inequality in the form

$$\begin{aligned} \langle u, [D, h'']u \rangle &\leqq 2c^{1/2} \| h^{1/2} u \| \, \| h'^{1/2} Du \| \\ &\leqq \langle Du, h'Du \rangle + c \langle u, hu \rangle, \end{aligned}$$

we obtain the identity

$$\partial_t \langle u, hu \rangle + 3 \langle Du, h'Du \rangle = \langle u, h'''u \rangle + 2 \langle u, hf \rangle \tag{2.44}$$

and from that estimate

$$\partial_t \langle u, hu \rangle + 2 \langle Du, h'Du \rangle \leqq c \langle u, hu \rangle + 2 |\langle u, hf \rangle|. \tag{2.45}$$

Omitting the term with Du, we obtain

$$\partial_t \| h^{1/2} u \|_2 \leqq (c/2) \| h^{1/2} u \|_2 + \| h^{1/2} f \|_2 \tag{2.46}$$

and by integration

$$\| h^{1/2} u(t) \|_2 \leqq e^{ct/2} \| h^{1/2} u_0 \|_2 + g(t), \tag{2.47}$$

from which we get (2.38) and (2.39) by taking successively $f = 0$ and $u_0 = 0$. Inequality (2.40) follows immediately from (2.39) and the definition of g. Next we integrate (2.45) in the interval $[0, T]$ to obtain

$$\begin{aligned} &\| h^{1/2} u(T) \|_2^2 + 2 \| h'^{1/2} Du;\, L^2([0, T], L^2) \|^2 \\ &\quad \leqq \| h^{1/2} u_0 \|_2^2 + \int_0^T dt\, \{ c \| h^{1/2} u(t) \|_2^2 + 2 \| h^{1/2} u(t) \|_2 \| h^{1/2} f(t) \|_2 \}. \end{aligned} \tag{2.48}$$

We omit the first term in the left-hand side of (2.48) and apply the resulting inequality successively with $f = 0$ and with $u_0 = 0$. For $f = 0$, we estimate the right-hand side by substituting (2.47) with $g = 0$ and obtain

$$\cdots \leqq \| h^{1/2} u_0 \|_2^2 \left[1 + c \int_0^T dt\, e^{ct} \right] = e^{cT} \| h^{1/2} u_0 \|_2^2. \tag{2.49}$$

This proves (2.41). For $u_0 = 0$, we estimate the right-hand side of (2.48) by substituting (2.47) with $u_0 = 0$ and obtain

$$\cdots \leqq \int_0^T dt\, \{ c g(t)^2 + 2 g(t) \| h^{1/2} f(t) \|_2 \} = \int_0^T dt\, 2 g(t) g'(t) = g(T)^2 \tag{2.50}$$

after noting that

$$g'(t) = \| h^{1/2} f(t) \|_2 + (c/2) g(t).$$

Formula (2.42) then follows from (2.48) continued by (2.50). □

We finally obtain a set of useful estimates by interpolation between those of Lemma 2.4 and those of Lemma 2.5.

LEMMA 2.6. *Let h_0 be a continuous function satisfying* (2.27) *and let h_1 satisfy the assumptions made on h in Lemma* 2.5. *Then the following estimate holds*:

$$\left\| h_0^{1/2-1/r} h_1^{1/r} \int_0^{\cdot} d\tau\, U(\cdot-\tau) Df(\tau);\, L^q([0,T],L^r) \right\| \leqq C(T) \| h_0^{1/2-1/r} h_1^{1/r} (1+x_+)^{1/4-(1/2r)} f;\, L^l([0,T],L^r) \| \tag{2.51}$$

for all $T>0$ and all r, $2 \leqq r \leqq \infty$, with

$$\frac{1}{r} < \frac{1}{q} = \frac{1}{l} - \frac{1}{4} - \frac{1}{2r} < \frac{3}{4} - \frac{1}{2r} \tag{2.52}$$

for $r>2$, and equality everywhere for $r=2$.

Proof. The result follows by interpolation between (2.36) written with h, q, r replaced by h_0, q_0, r_0, and (2.42) with h replaced by h_1. In particular q, l, and r are related to q_0, l_0, and r_0 by the convexity relations

$$\frac{1}{l} = \frac{1-2/r}{l_0} + \frac{2}{r}, \qquad \frac{1}{q} = \frac{1-2/r}{q_0} + \frac{1}{r}$$

and condition (2.52) simply expresses the condition $0 < 1/q_0 = 1/l_0 - \frac{1}{4} < \frac{3}{4}$ from Lemma 2.4. □

In the applications, we will often use smooth weighting factors that satisfy in particular the assumptions of Lemmas 2.4 and 2.5, decreasing exponentially when x tends to $-\infty$ and increasing as a power when x tends to $+\infty$. We will use a two-parameter family $\{h_{\alpha\beta}\}$, $\alpha>0$, $\beta \geqq 0$, satisfying the following properties: $h_{\alpha\beta} \in \mathscr{C}^\infty$, $h_{\alpha\beta}(x) = \exp(\alpha x)$ for $x \leqq 0$, $h_{\alpha\beta}(x) \sim Cx^\beta$ when $x \to +\infty$, $0 \leqq h'_{\alpha\beta} \leqq \alpha h_{\alpha\beta}$, and additional estimates for higher derivatives. For definiteness, we choose a nonnegative $\mathscr{C}^\infty$ function φ with compact support contained in $[-1,1]$ and with $\|\varphi\|_1 = 1$, and we define for $\alpha>0$, $\beta \geqq 0$

$$\omega_{\alpha\beta}(x) = \alpha \quad \text{for } x \leqq 1,$$
$$\omega_{\alpha\beta}(x) = \beta(x-1+\beta/\alpha)^{-1} \quad \text{for } x \geqq 1,$$
$$\tilde{\omega}_{\alpha\beta} = \omega_{\alpha\beta} * \varphi, \qquad h_{\alpha\beta}(x) = \exp\left\{ \int dy\, \tilde{\omega}_{\alpha\beta}(y) \right\}.$$

This family of functions meets all our requirements and will be used without further comments.

3. Uniqueness of L^2 solutions. In this section we derive our results on the uniqueness of L^2 solutions of the GKdV equation (1.1), more precisely of solutions in $(L^\infty_{\mathrm{loc}} \cap \mathscr{C}_W)([0,T],L^2)$ for some $T>0$, and we discuss the uniqueness classes that emerge naturally from them. The method is standard and consists of showing that a suitable norm of the difference of two solutions with the same initial data satisfies a linear inequality that compels it to be zero. Since the problem is local, it is sufficient to consider small time intervals, and since the equation is time-translation invariant, we can take the initial time to be zero. We first proceed formally and consider two solutions u_1, u_2 of (1.1) with common initial data $u_1(0) = u_2(0) = u_0$, defined in a time interval $[0,T)$. The difference $w = u_1 - u_2$ satisfies the equation

$$\partial_t w + D^3 w = DV'(u_1) - DV'(u_2) = D(\tilde{V}'' w) \tag{3.1}$$

where

$$\tilde{V}'' = \int_0^1 d\lambda\, V''(\lambda u_1 + (1-\lambda)u_2). \tag{3.2}$$

Since $w(0)=0$, w satisfies the integral equation

$$w(t)=\int_0^t d\tau\, U(t-\tau)D(\tilde{V}''w(\tau)). \tag{3.3}$$

Under suitable mild assumptions on V, u_1, and u_2, (3.1) and (3.3) make sense and are equivalent (under the initial condition $w(0)=0$) in a weak (distributional) sense. We will often assume that V satisfies the condition (1.5) or equivalently

$$|V''(\rho)|\leqq C|\rho|^p \tag{3.4}$$

for some $p>0$ and all $\rho\in\mathbb{R}$. We consider successively the cases $p\leqq 2$ and $p\geqq 2$. Our basic uniqueness result in the case $p\leqq 2$ is the following proposition.

PROPOSITION 3.1. *Let V satisfy* (3.4) *with $0<p\leqq 2$, let $u_0\in L^2$, and let $T>0$ (possibly $T=\infty$). Then* (1.1) *with initial data $u(0)=u_0$ has at most one solution satisfying*

$$u\in(L^\infty_{\mathrm{loc}}\cap\mathscr{C}_w)([0,T),L^2), \tag{3.5}$$

$$(1+x_+)^{\beta/2}u\in L^{q_1}_{\mathrm{loc}}([0,T),L^2), \tag{3.6}$$

$$h_{\alpha 0}^{1/2}u\in L^{q_1}_{\mathrm{loc}}([0,T),L^r)\quad \textit{for some } \alpha>0 \tag{3.7}$$

where

$$\frac{2}{q_1}=\beta=\frac{1}{p}-\frac{1}{4}, \tag{3.8}$$

$$\frac{1}{r}=\frac{1}{2}-\frac{p}{4}, \tag{3.9}$$

$$\frac{1}{q}<\frac{3}{4}-\frac{1}{2r}\equiv\frac{1}{2}+\frac{p}{8}. \tag{3.10}$$

Proof. Let u_1 and u_2 be two solutions satisfying (3.5)–(3.7) with common initial data $u_1(0)=u_2(0)=u_0$ and let $h=h_{\alpha 0}$. By (3.7), $h^{1/2}w\in L^q_{\mathrm{loc}}([0,T),L^r)$. Without loss of generality, we can assume that $q<r$. We estimate $h^{1/2}w$ by Lemma 2.6 with $h_1=h_{\alpha 1}$ and $h_0=h=h_{\alpha 0}$ and obtain for any $t\in[0,T)$

$$\|h^{1/2}w;\, L^q([0,T],L^r)\|\leqq C\|h^{1/2}(1+x_+)^{1/4+1/(2r)}\tilde{V}''w;\, L^l([0,t],L^{\bar{r}})\| \tag{3.11}$$

with q, r, and l satisfying (2.52). We next estimate the right-hand side by the Hölder inequality as

$$\cdots\leqq C\|h^{1/2}w;\, L^q([0,t],L^r)\|\,\|(1+x_+)^{1/4+1/(2r)}\tilde{V}'';\, L^m([0,t],L^s)\| \tag{3.12}$$

with $1/s=1-2/r$ and $1/m=1/l-1/q=\frac{1}{4}+1/(2r)$. The last norm is then estimated by using (3.6) as

$$\cdots\leqq C\sum_{i=1,2}\|(1+x_+)^{\beta/2}u_i;\, L^{q_1}([0,t],L^{r_1})\|^p \tag{3.13}$$

with

$$\beta=\frac{2}{q_1}=\frac{1}{2p}+\frac{1}{rp},\qquad \frac{1}{r_1}=\frac{1}{ps}=\frac{1}{p}-\frac{2}{rp}, \tag{3.14}$$

or equivalently

$$\beta=\frac{2}{q_1}=\frac{1}{p}-\frac{1}{2r_1},\qquad \frac{1}{r}=\frac{1}{2}-\frac{p}{2r_1}. \tag{3.15}$$

For $0<p\leqq 2$, we can choose $r_1=2$, so that (3.15) reduces to (3.8), (3.9), while the second inequality in (2.52) reduces to (3.10). Recalling (3.11)–(3.13), using the assumption (3.6), and taking t sufficiently small, we obtain $h^{1/2}w=0$ and therefore $w=0$ in $[0, t]$. □

Remark 3.1. For low values of p, (3.8) yields unpleasant values $q_1<1$. This is seen to arise from a natural assumption on $|u|^p$ and to create no trouble.

Remark 3.2. The values of q, r, and q_1 ensure in particular that $|u|^{p+1}\in L^l(L^{\bar r})$ with $l>1$, $\bar r\geqq 1$, so that the equation (1.1) makes sense in $\mathcal{D}'$.

Next we turn to the case $p\geqq 2$. The basic uniqueness result follows.

PROPOSITION 3.2. *Let V satisfy* (3.4) *with $p\geqq 2$, let $u_0\in L^2$, and let $T>0$ (possibly $T=\infty$). Then* (1.1) *with initial data $u(0)=u_0$ has at most one solution satisfying* (3.5) *and*

$$(3.16)_+ \qquad \chi_+(1+x)^{1/4}|u|^p\in L^{q_1}_{\mathrm{loc}}([0, T), L^1) \quad \textit{for some } q_1>4,$$

$$(3.16)_- \qquad \chi_-|u|^p\in l^\infty(L^{q_1}_{\mathrm{loc}}([0, T), L^1)) \quad \textit{for some } q_1>4,$$

$$(3.17)_+ \qquad \chi_+u\in L^q_{\mathrm{loc}}([0, T), L^\infty) \quad \textit{for some } q>\tfrac{4}{3},$$

$$(3.17)_- \qquad \chi_-\, e^{\alpha x/2}u\in l^\infty(L^q_{\mathrm{loc}}([0, T), L^\infty)) \quad \textit{for some } q>\tfrac{4}{3} \textit{ and some } \alpha>0.$$

Proof. Let u_1 and u_2 be two solutions satisfying (3.5), $(3.16)_\pm$, and $(3.17)_\pm$ with initial data $u_1(0)=u_2(0)=u_0$. Let $h=h_{\alpha 0}$. By $(3.17)_\pm$, $\chi_+h^{1/2}w\in L^q_{\mathrm{loc}}([0, T), L^\infty)$ and $\chi_+h^{1/2}w\in l^\infty(L^q_{\mathrm{loc}}([0, T), L^\infty))$. We estimate $h^{1/2}w$ by inserting the estimate (2.28) of Lemma 2.3 into the integral equation (3.3) and obtain

$$\begin{aligned}|h^{1/2}w(t, x)|\leqq C(t)\int_0^t d\tau\,|t-\tau|^{-3/4}\\ \cdot\Big\{\int_{y\geqq 0} dy\,(1+y)^{1/4}|\tilde V''(\tau, y)||h^{1/2}w(\tau, y)|\\ +\int_{y\leqq 0} dy\,\exp(-\gamma|x-y|/2)|\tilde V''(\tau, y)||h^{1/2}w(\tau, y)|\Big\}\\ \equiv J_+(x)+J_-(x).\end{aligned} \tag{3.18}$$

The contribution $J_+(x)$ of the region $y\geqq 0$ is estimated in $L^q([0, t])$ by the Hardy–Littlewood–Sobolev inequality [8, p. 117] as

$$\begin{aligned}\|J_+;\, L^q([0, t], L^\infty)\|\leqq C(t)\|\chi_+h^{1/2}w;\, L^q([0, t], L^\infty)\|\\ \cdot\|\chi_+(1+\cdot)^{1/4}\tilde V'';\, L^4([0, t], L^1)\|\end{aligned} \tag{3.19}$$

and the last norm is estimated by the use of (3.4) and $(3.16)_+$ as

$$C\sum_{i=1,2} t^{1/4-1/q_1}\|\chi_+(1+\cdot)^{1/4}|u_i|^p;\, L^{q_1}([0, t], L^1)\|. \tag{3.20}$$

The contribution $J_-(x)$ of the region $y\leqq 0$ is decomposed into the sum of the contributions of unit intervals:

$$J_-(x)\leqq C(t)\sum_{j\in\mathbf{Z}^-}\exp(-\gamma|x-j|/2)\int_0^t d\tau\,|t-\tau|^{-3/4}\|\chi_j\chi_-h^{1/2}w(\tau)\|_\infty\|\chi_j\chi_-\tilde V''(\tau)\|_1.$$

For $x \geqq 0$ we omit x in the exponential and estimate the time integral as before to obtain

$$\begin{aligned} \|\chi_+ J_-; L^q([0,t], L^\infty)\| &\leqq C(t) \sum_{j \leqq 0} e^{\gamma j/2} \|\chi_j \chi_- h^{1/2} w; L^q([0,t], L^\infty)\| \\ &\quad \cdot \|\chi_j \chi_- \tilde{V}''; L^4([0,t], L^1)\| \\ &\leqq C(t) t^{1/4-1/q_1} \|\chi_- h^{1/2} w; l^\infty(L^q([0,t], L^\infty))\| \\ &\quad \cdot \sum_{i=1,2} \|\chi_- |u_i|^p; l^\infty(L^{q_1}([0,t], L^1))\| \end{aligned} \tag{3.21}$$

by taking the sup over j of the last two norms in the middle member and using $(3.16)_-$. For $x \leqq 0$, we estimate similarly (with $k \leqq 0$)

$$\begin{aligned} \|\chi_k J_-; L^q([0,t], L^\infty)\| &\leqq C(t) \sum_{j \leqq 0} \exp(-\gamma|j-k|/2) \\ &\quad \cdot \|\chi_j \chi_- h^{1/2} w; L^q([0,t], L^\infty)\| \, \|\chi_j \chi_- \tilde{V}''; L^4([0,t], L^1)\| \end{aligned}$$

so that by the Young inequality in l' spaces and $(3.16)_-$ again

$$\begin{aligned} \|\chi_- J_-; l^\infty(L^q([0,t], L^\infty))\| &\leqq C(t) t^{1/4-1/q_1} \|\chi_- h^{1/2} w; l^\infty(L^q([0,t], L^\infty))\| \\ &\quad \times \sum_{i=1,2} \|\chi_- |u_i|^p; l^\infty(L^{q_1}([0,t], L^1))\|. \end{aligned} \tag{3.22}$$

Collecting the estimates (3.19)-(3.22) and defining

$$|||w||| = \|\chi_+ h^{1/2} w; L^q([0,t], L^\infty)\| + \|\chi_- h^{1/2} w; l^\infty(L^q([0,t], L^\infty))\|, \tag{3.23}$$

we obtain a linear inequality

$$|||w||| \leqq C(t) t^{1/4-1/q_1} M |||w||| \tag{3.24}$$

with

$$M = \sum_{i=1,2} \{\|\chi_+ (1+\cdot)^{1/4} |u_i|^p; L^{q_1}([0,t], L^1)\| + \|\chi_- \gamma u_i|^p; l^\infty(L^{q_1}([0,t], L^1))\|\}. \tag{3.25}$$

It follows from (3.24) by taking t sufficiently small that $|||w||| = 0$ and therefore that $w = 0$ in $[0, t]$. □

Remark 3.3. We could have stated $(3.16)_\pm$ in a slightly weaker form by taking $q_1 = 4$ and assuming in addition that for any fixed $s \in [0, T)$

$$\lim_{t > s, t \downarrow s} \|\chi_- |u|^p; l^\infty(L^4([s,t], L^1))\| = 0.$$

However, the only practical way to enforce that condition is through $(3.16)_-$ as stated with $q_1 > 4$. The condition $q_1 > 4$ is not really needed in $(3.16)_+$, where $q_1 = 4$ would be sufficient, and has been imposed for consistency.

We now turn to analyzing (3.5)-(3.7) and (3.5), $(3.16)_\pm$, $(3.17)_\pm$ that appear in Propositions 3.1 and 3.2, respectively, and in particular to deriving a priori estimates of the corresponding norms in terms of the initial data for smooth solutions of the equation (1.1). The degree of smoothness required in the following arguments is actually rather low. Aside from the decay at $+\infty$ in space that occurs, for instance, in (3.6) and $(3.16)_+$ and will have to be assumed explicitly in some form, it will be sufficient that the solutions satisfy $u \in (L^\infty_{\mathrm{loc}} \cap \mathscr{C}_w)([0, T), H^1)$ for all the other conditions to be satisfied and all subsequent calculations to make sense (see in particular Remark 3.7 below).

The first task is to dispose of the harmless assumptions (3.7) in Proposition 3.1 and $(3.17)_\pm$ in Proposition 3.2. This is done in Propositions 3.3 and 3.4 below.

PROPOSITION 3.3. *Let V satisfy*

$$|V'(\rho)| \leqq C|\rho|^{p+1} \tag{3.26}$$

with $0<p\leqq 2$, *let* $u_0\in L^2$, *let* $T>0$ (*possibly* $T=\infty$), *and let* u *be a solution of* (1.1) *with initial data* $u(0)=u_0$ *satisfying* (3.5)–(3.7) *with* (3.9), (3.10), *and*

$$\frac{2}{q_1}<\beta=\frac{1}{p}-\frac{1}{4}, \tag{3.27}$$

$$\max\left(\frac{1}{2}-\frac{p}{4},\frac{p}{12}\right)\equiv\max\left(\frac{1}{r},\frac{1}{6}-\frac{1}{3r}\right)<\frac{1}{q}. \tag{3.28}$$

Then, for any $t\in[0,T)$ *the norm of* $h_{\alpha 0}^{1/2}u$ *in* $L^q([0,t],L^r)$ *is estimated a priori in terms of* t, *of the norm of* u *in* $L^\infty([0,t],L^2)$ *and of the norm of* $(1+x_+)^{\beta/2}u$ *in* $L^{q_1}([0,t],L^2)$.

Proof. The proof is a slight variant of that of Proposition 3.1. We prove the result in successive intervals $[s_j,s_{j+1}]$ covering the interval $[0,t]$, with $s_0=0$. In each such interval, u satisfies the integral equation

$$u(t)=U(t-s)u(s)+\int_s^t d\tau\, U(t-\tau)DV'(u(\tau)) \tag{3.29}$$

with $s=s_j$. The free term in (3.29) is estimated by the use of (2.12) and the Hölder inequality as

$$\|h_{\alpha 0}^{1/2}U(\cdot-s)u(s);L^q(I,L^r)\|\leqq C|I|^{1/q-1/6+(1/3r)}\|u(s)\|_2 \tag{3.30}$$

with $I=[s_j,s_{j+1}]$. The integral in (3.29) is estimated in $L^q(I,L^r)$ by the same method as in the proof of Proposition 3.1 as

$$C\|h_{\alpha 0}^{1/2}u;L^q(I,L^r)\|\,\|(1+x_+)^{\beta/2}u;L^{q_1}(I,L^{r_1})\|^p|I|^\varepsilon \tag{3.31}$$

with $\varepsilon=\frac{1}{2}-p/8-p/q_1>0$. Now the right-hand side of (3.30) is obviously estimated in terms of the norm of u in $L^\infty(\cdot,L^2)$, while the coefficient of the first norm in (3.31) can be made less than or equal to $\frac{1}{2}$ (for instance) by taking I sufficiently small, depending only on the norm of $(1+x_+)^{\beta/2}u$ in $L^{q_1}(\cdot,L^{r_1})$, thereby providing an a priori estimate of the norm of $h_{\alpha 0}^{1/2}u$ in $L^q(I,L^r)$ in terms of the relevant quantities. □

PROPOSITION 3.4. *Let* V *satisfy* (3.26) *with* $p\geqq 2$, *let* $u_0\in L^2$, *let* $T>0$ (*possibly* $T=\infty$), *and let* u *be a solution of* (1.1) *with initial data* $u(0)=u_0$ *satisfying* (3.5), $(3.16)_\pm$, *and* $(3.17)_\pm$ *with* $\frac{4}{3}<q<6$.

Then, for any $t\in[0,T)$, *the norms of* χ_+u *in* $L^q([0,t],L^\infty)$ *and of* $\chi_-h_{\alpha 0}^{1/2}u$ *in* $l^\infty(L^q([0,t],L^\infty))$ *are estimated a priori in terms of* t *and of the norms of* u *in* $L^\infty([0,t],L^2)$, *of* $\chi_+(1+x)^{1/4}|u|^p$ *in* $L^{q_1}([0,t],L^1)$, *and of* $\chi_-|u|^p$ *in* $l^\infty(L^{q_1}([0,t],L^1))$.

Proof. The proof proceeds along the same lines as that of Proposition 3.3 by using the estimates of the proof of Proposition 3.2 instead of those of the proof of Proposition 3.1. We again prove the result in successive intervals $[s_j,s_{j+1}]$ covering $[0,t]$, with $s_0=0$. In each such interval I, u satisfies the integral equation (3.29) with $s=s_j$. The free term is now estimated by the use of (2.14) and with $h=h_{\alpha 0}$ as

$$\|h^{1/2}U(\cdot-s)u(s);L^q(I,L^\infty)\|\leqq C|I|^{1/q-1/6}\|u(s)\|_2. \tag{3.32}$$

The integral in (3.29) is estimated by the use of (2.28) and (3.26) as

$$\begin{aligned} h^{1/2}\int_s^t d\tau\, U(t-\tau)DV'(u(\tau))&\leqq C(|I|)\int_s^t d\tau\,|t-\tau|^{-3/4}\\ &\cdot\left\{\int_{y\geqq 0}dy\,(1+y)^{1/4}h^{1/2}|u(\tau,y)|^{p+1}\right.\\ &\left.+\int_{y\leqq 0}dy\exp(-\gamma|x-y|/2)h^{1/2}|u(\tau,y)|^{p+1}\right\}\\ &\equiv J_+(x)+J_-(x). \end{aligned} \tag{3.33}$$

We estimate J_+ and J_- as in the proof of Proposition 3.2 and obtain

$$\begin{aligned}(3.34)\qquad &\|J_+;\, L^q(I, L^\infty)\| \\ &\leqq C(|I|)|I|^{1/4-1/q_1}\|\chi_+ h^{1/2}u;\, L^q(I, L^\infty)\|\, \|\chi_+(1+\cdot)^{1/4}|u|^p;\, L^{q_1}(I, L^1)\|,\end{aligned}$$

$$\begin{aligned}(3.35)\qquad &\|\chi_+ J_-;L^q(I, L^\infty)\| + \|\chi_- J_-;\, l^\infty(L^q(I, L^\infty))\| \\ &\leqq C(|I|)|I|^{1/4-1/q_1}\|\chi_- h^{1/2}u;\, l^\infty(L^q(I, L^\infty))\|\, \|\chi_-|u|^p;\, l^\infty(L^{q_1}(I, L^1))\|.\end{aligned}$$

From (3.32), (3.33), and (3.35), we obtain an inequality of the form

$$(3.36)\qquad |||u||| \leqq C|I|^{1/q-1/6}\|u:\, L^\infty(I, L^2)\| + C(|I|)|I|^{1/4-1/q_1}M|||u|||$$

where $|||u|||$ is defined as in (3.23) and M as in (3.25), but with one single function u. The coefficient of $|||u|||$ in the right-hand side of (3.36) can be made less than or equal to $\frac{1}{2}$ (for instance) by taking I sufficiently small, depending only on the norms of u contained in M that correspond to $(3.16)_\pm$, thereby providing an a priori estimate of $|||u|||$ in terms of the latter norms. □

We now turn to the analysis of the main conditions (3.5), (3.6) of Proposition 3.1 and (3.5), $(3.16)_\pm$ of Proposition 3.2. An essential tool will be the following identity [10], [12] satisfied by the solutions of (1.1), a preliminary version of which has already been used in § 2:

$$(3.37)\qquad \partial_t\langle u, hu\rangle + 3\langle Du, h'Du\rangle = \langle u, h'''u\rangle - 2\int h'(uV'(u) - V(u))$$

or in integral form

$$\begin{aligned}(I.1)\qquad &\langle u, hu\rangle(t) + 3\int_0^t d\tau\, \langle Du, h'Du\rangle(\tau) \\ &= \langle u_0, hu_0\rangle + \int_0^t d\tau \left\{\langle u, h'''u\rangle(\tau) - 2\int h'(uV'(u) - V(u))(\tau)\right\}.\end{aligned}$$

We will henceforth refer to this identity as (I.1). Clearly it makes sense, for all $t \in [0, T)$, for $h \in \mathscr{C}^3(\mathbb{R}, \mathbb{R})$ with compactly supported h', provided

$$(3.38)\qquad u \in (L^\infty_{\mathrm{loc}} \cap \mathscr{C}_w)([0, T), L^2) \cap L^2_{\mathrm{loc}}([0, T), H^1_{\mathrm{loc}}),$$

$$(3.39)\qquad uV'(u) - V(u) \in L^1_{\mathrm{loc}}([0, T), L^1_{\mathrm{loc}}).$$

It is known that under suitable assumptions, (I.1) for positive increasing h implies that $h^{1/2}u \in L^\infty_{\mathrm{loc}}(L^2)$ as soon as $h^{1/2}u_0 \in L^2$ [10], [12]. Since we will make repeated use of that fact in various contexts, we prove it under assumptions that seem reasonably optimal. For that purpose we need some auxiliary estimates, which will arise as part of Lemma 3.2 below. We state that lemma here for convenience, although we will not need it in its full generality until the proof of Proposition 4.4 below. We begin with an elementary inequality that we give without proof.

LEMMA 3.1. *Let $0 \leqq \delta < 1$. Then for any $\varepsilon > 0$ and any $x, y \in \mathbb{R}^+$, the following estimate holds:*

$$(3.40)\qquad x^\delta y^{1-\delta} \leqq \varepsilon x + b_\varepsilon y$$

where

$$b_\varepsilon = (1-\delta)(\delta/\varepsilon)^{\delta/(1-\delta)}.$$

LEMMA 3.2. *Let $W \in \mathscr{C}(\mathbb{R}, \mathbb{R}^+)$ satisfy*

$$0 \leqq W(\rho) \leqq \varepsilon\rho^6 + a_\varepsilon|\rho|^{p+2}$$

for some p, $0 \leqq p < 4$ and all $\rho \in \mathbb{R}$. Let $h \in \mathscr{C}^1(\mathbb{R}, \mathbb{R}^+)$, $h > 0$. Then for any $\varepsilon' > 0$, the following estimate holds:

$$\begin{aligned}\|hW(u)\|_1 \leqq 5\{&(\varepsilon\|u\|_2^4 + \varepsilon' a_\varepsilon \|u\|_2^p)\langle Du, hDu\rangle + \varepsilon\|u\|_2^4\langle u, h^{-1}h'^2u\rangle \\ &+ a_\varepsilon\langle u, h^\nu u\rangle^{1+p/4}\langle u, h^{-1}h'^2u\rangle^{p/4} + a_\varepsilon b_{\varepsilon'}\langle u, h^\nu u\rangle^{1/\nu}\|u\|_2^{p+2-2/\nu}\}\end{aligned} \tag{3.41}$$

with $\nu = (4-p)/(4+p)$, for all $u \in L^2 \cap H^1_{\mathrm{loc}}$ for which all norms in the right-hand side are finite.

Proof. Clearly

$$\|hW(u)\|_1 \leqq \varepsilon I_4 + a_\varepsilon I_p \tag{3.42}$$

where

$$\begin{aligned}I_p = \||h|u|^{p+2}\|_1 &\leqq \langle u, h^\nu u\rangle \|h^{1-\nu}|u|^p\|_\infty \\ &= \langle u, h^\nu u\rangle \|h^{(1+\nu)/2}u^2\|_\infty^{p/2}.\end{aligned}$$

Now by an elementary computation

$$\begin{aligned}|h^{(1+\nu)/2}u^2| &\leqq [(1+\nu)/2]\langle u, h^{(\nu-1)/2}|h'|u\rangle + 2\|h^{\nu/2}u\|_2\|h^{1/2}Du\|_2 \\ &\leqq 5^{1/2}\langle u, h^\nu u\rangle^{1/2}\{\langle u, h^{-1}h'^2u\rangle + \langle Du, hDu\rangle\}^{1/2}\end{aligned} \tag{3.43}$$

by the Schwarz inequality, so that

$$I_p \leqq 5\langle u, h^\nu u\rangle^{1+p/4}\{\langle u, h^{-1}h'^2u\rangle^{p/4} + \langle Du, hDu\rangle^{p/4}\}. \tag{3.44}$$

We use (3.44) directly with p replaced by 4 to estimate I_4. On the other hand, we estimate the term containing Du in I_p by applying Lemma 3.1 with $\delta = p/4$ and $x = \langle Du, hDu\rangle\|u\|_2^p$, and obtain

$$\langle u, h^\nu u\rangle^{1+p/4}\langle Du, hDu\rangle^{p/4} \leqq \varepsilon'\|u\|_2^p\langle Du, hDu\rangle + b_{\varepsilon'}\langle u, h^\nu u\rangle^{1/\nu}\|u\|_2^{p+2-2/\nu}. \tag{3.45}$$

Substituting (3.45) into (3.44) and then into (3.42) yields (3.41). □

We now prove the basic set of estimates that can be derived from the identity (I.1).

PROPOSITION 3.5. *Let V satisfy the condition*

$$\lim_{|\rho|\to\infty} |\rho|^{-6}(\rho V'(\rho) - V(\rho))_- = 0. \tag{3.46}$$

Let $T > 0$, let $u_0 \in L^2$, and let u be a solution of the equation (1.1) *with initial data $u(0) = u_0$, satisfying* (3.38) *and* (3.39), *and such that* (I.1) *holds for all $h \in \mathscr{C}^3$ with compactly supported h'. Then*:

(1) *$Du \in l^\infty(L^2_{\mathrm{loc}}([0, T), L^2))$ and for all $t \in [0, T)$, Du is estimated in $l^\infty(L^2([0, t], L^2))$ in terms of t and of $\|u_0\|_2$.*

(2) *Let $h \in \mathscr{C}^3(\mathbb{R}, \mathbb{R}^+)$ satisfy $0 \leqq h' \leqq c_1(1+h)$, $h''^2 \leqq c_2 h'(1+h)$, $h''' \leqq c_3(1+h)$, and let $h^{1/2}u_0 \in L^2$. Then $h^{1/2}u \in L^\infty_{\mathrm{loc}}([0, T), L^2)$, $h'^{1/2}Du \in L^2_{\mathrm{loc}}([0, T), L^2)$ and for all $t \in [0, T)$, $h^{1/2}u$ is estimated in $L^\infty([0, t], L^2)$ and $h'^{1/2}Du$ is estimated in $L^2([0, t], L^2)$ in terms of t and of $\|(1+h)^{1/2}u_0\|_2$. In addition,* (I.1) *holds for that specific h.*

Remark 3.4. Condition (3.46) is satisfied, in particular, if V satisfies either (3.4) or (3.26) for some p, $0 \leqq p < 4$. Furthermore, in that case the integrability condition (3.39) follows from (3.38).

Proof of Proposition 3.5. We first note that (I.1) with $h \equiv 1$ implies the conservation of the L^2 norm, namely, $\|u(t)\|_2 = \|u_0\|_2$ for all $t \in [0, T)$. Condition (3.46) is not needed for that remark.

Next we consider $h \in \mathscr{C}^3(\mathbb{R}, \mathbb{R}^+)$ with compactly supported h'. From (3.46) and the conditions $V \in \mathscr{C}^2$, $V(0) = V'(0) = 0$, it follows that for any $\varepsilon > 0$, there exists $a_\varepsilon \geqq 0$ such that

$$(\rho V'(\rho) - V(\rho))_- \leqq \varepsilon \rho^6 + a_\varepsilon \rho^2 \tag{3.47}$$

for all $\rho \in \mathbb{R}$. We now estimate the last integral in the right-hand side of (I.1) by using (3.47), estimating the contribution of the term $\varepsilon\rho^6$ in the same way as in the proof of Lemma 3.2, with h replaced by h' (see especially (3.44) with $p = 4$, $\nu = 0$). We obtain

$$-\int h'(uV'(u) - V(u)) \leqq 5\varepsilon \|u\|_2^4 \langle Du, h'Du\rangle + (5\varepsilon c_2 \|u\|_2^4 + a_\varepsilon c_1)\langle u, (1+h)u\rangle. \tag{3.48}$$

Taking ε sufficiently small, depending only on $\|u_0\|_2$, for instance,

$$\varepsilon \|u\|_2^4 = \tfrac{1}{10},$$

we obtain from (I.1)

$$\langle u, hu\rangle(t) + 2\int_0^t d\tau \langle Du, h'Du\rangle(\tau) \leqq \langle u_0, hu_0\rangle + C\int_0^t d\tau \langle u, (1+h)u\rangle(\tau) \tag{3.49}$$

where C depends on $\|u_0\|_2$ and on h through c_1, c_2, and c_3 only.

We now prove part 1. For that purpose, we choose a function $\psi_0 \in \mathscr{C}^3$ with $0 \leqq \psi_0 \leqq 1$, $\psi_0(x) = 1$ (respectively, 0) for $|x| \leqq \frac{1}{2}$ (respectively, $|x| \geqq 1$), we define ψ_j by $\psi_j(x) = \psi_0(x - j)$ for all $j \in \mathbb{Z}$, and we apply (3.49) with $h = h_j$ defined by

$$h_j(x) = \int_{-\infty}^x \psi_j(y)\, dy,$$

thereby obtaining

$$2\|\chi_j Du;\, L^2([0, t], L^2)\|^2 \leqq (1 + Ct)(1 + \|\psi_0\|_1)\|u_0\|_2^2.$$

This proves part 1.

To prove part 2 under the relevant assumptions on h, we approximate h by a sequence h_N with compactly supported h'_N such that $h_N(x) = h(x)$ for $|x| \leqq N$, $h_N(x) =$ const. for $x \geqq N+1$ and for $x \leqq -(N+1)$, and such that h_N satisfies the same assumptions as h with constants c_1, c_2, c_3 uniform in N. We then apply (3.49) with h replaced by h_N and obtain from it estimates for $\langle u, h_N u\rangle$ and for $\langle Du, h'_N Du\rangle$ in terms of $\langle u_0, (1+h)u_0\rangle$, uniformly in N, by the Gronwall inequality. Taking the limit $N \to \infty$ yields the required estimates with the given h. Finally, (I.1) for that h follows from the estimates and from the Dominated Convergence Theorem. □

We now return to the analysis of conditions (3.5), (3.6) of Proposition 3.1 and $(3.16)_\pm$, $(3.17)_\pm$ of Proposition 3.2. In both cases, the conservation of the L^2-norm for (smooth) solutions satisfying (I.1) yields an a priori estimate in $L^\infty([0, T), L^2)$, thereby disposing of (3.5).

Next we consider condition (3.6) of Proposition 3.1 corresponding to $p \leqq 2$ in (3.4). In that case, it follows from Proposition 3.5 applied with $h = h_{0\beta}$ that smooth solutions of the equation (1.1) satisfy an a priori estimate in the space defined by (3.6), actually with $q_1 = \infty$, in terms of $\|(1 + x_+)^{\beta/2} u_0\|_2$. That result, together with Proposition 3.3 and L^2-norm conservation indicates that for $p \leqq 2$, (3.5)–(3.7) with (3.9), (3.10), (3.27), (3.28) define a suitable class for existence and uniqueness, with u_0 satisfying only the assumption $(1 + x_+)^{\beta/2} u_0 \in L^2$ with $\beta = 1/p - \frac{1}{4}$.

We now turn to the analysis of the main condition $(3.16)_{\pm}$ that occurs in Proposition 3.2 corresponding to the more difficult case $p \geqq 2$. We first give sufficient conditions for the corresponding norms of smooth solutions to satisfy a priori estimates. That can be done in either of two ways, the first of which is based on Proposition 3.5. We will need two auxiliary results. The first is the following elementary Sobolev estimate, which we give without proof.

LEMMA 3.3. *Let I be a bounded interval, let χ be its characteristic function, and let $u \in H^1(I)$. Then the following estimate holds*:

$$\|\chi u\|_\infty^2 \leqq |I|^{-1}\|u\|_2^2 + 2\|\chi u\|_2\|\chi Du\|_2. \tag{3.50}$$

The second auxiliary result that we need is the following lemma.

LEMMA 3.4. *Let $h \in \mathscr{C}(\mathbb{R}, \mathbb{R}^+)$ satisfy $0 \leqq h' \leqq c_1 h$ and $|h''| \leqq c_2 (hh')^{1/2}$. Let $t > 0$, and let u be such that $h^{1/2}u \in L^\infty([0, t], L^2)$ and $h'^{1/2}Du \in L^2([0, t], L^2)$. Let $2 \leqq r < \infty$, and let $h_1 \in \mathscr{C}(\mathbb{R}, \mathbb{R}^+)$ satisfy $h_1^2 \leqq h^{r/2+1}h'^{r/2-1}$. Then $h_1|u|^2 \in L^{q_1}([0, t], L^1)$ and the following estimate holds*:

$$\begin{aligned} &\|h_1|u|^r; L^{q_1}([0, t], L^1)\| \\ &\quad \leqq C\{t^{1/q_1}\|h^{1/2}u; L^\infty([0, t], L^2)\|^r + \|h^{1/2}u; L^\infty([0, t], L^2)\|^{r/2+1} \\ &\qquad \cdot \|h'^{1/2}Du; L^2([0, t], L^2)\|^{r/2-1}\} \end{aligned} \tag{3.51}$$

with $1/q_1 = r/4 - \frac{1}{2}$.

Proof. We estimate pointwise for $\tau \in [0, t]$

$$\|h_1|u|^r\|_1 \leqq \|h^{1/2}u\|_2^2 \|h_2 u^2\|_\infty^{r/2-1} \tag{3.52}$$

with $h_2 = (hh')^{1/2}$. Then by derivation and integration

$$\|h_2 u^2\|_\infty \leqq \tfrac{1}{2}\langle u, |h_2'|u\rangle + \|h^{1/2}u\|_2 \|h'^{1/2}Du\|_2. \tag{3.53}$$

Now

$$|h_2'| = \tfrac{1}{2}(hh')^{-1/2}|h'^2 + hh''| \leqq \tfrac{1}{2}(c_1^{3/2} + c_2)h$$

so that

$$\|h_1|u|^r\|_1 \leqq C\{\|h^{1/2}u\|_2^r + \|h^{1/2}u\|_2^{r/2+1}\|h'^{1/2}Du\|_2^{r/2-1}\} \tag{3.54}$$

from which the result follows by taking the L^{q_1} norm and applying the Hölder inequality in time. □

We can now derive a first set of a priori estimates of the norms that correspond to $(3.16)_{\pm}$ for sufficiently smooth solutions of (1.1) with V satisfying (3.4), $2 \leqq p < 3$, and exhibit a new uniqueness class in that case.

PROPOSITION 3.6. (1) *Let $2 \leqq p < 3$ and $\beta = \frac{1}{2} - 1/(2p)$. Let $T > 0$, and let us satisfy*

$$(1 + x_+)^{\beta/2}u \in L^\infty_{\mathrm{loc}}([0, T), L^2), \tag{3.55}$$

$$\chi_+(1 + x)^{(\beta-1)/2}Du \in L^2_{\mathrm{loc}}([0, T), L^2), \tag{3.56}_+$$

$$\chi_- Du \in l^\infty(L^2_{\mathrm{loc}}([0, T), L^2)). \tag{3.56}_-$$

Then $(3.16)_{\pm}$ is satisfied and the following estimates hold:

$$\begin{aligned} &\|\chi_+(1 + x)^{1/4}|u|^p; L^{q_1}([0, t], L^1)\| \\ &\quad \leqq C\{t^{1/q_1}\|h_{\alpha\beta}^{1/2}u; L^\infty([0, t], L^2)\|^p + \|h_{\alpha\beta}^{1/2}u; L^\infty([0, t], L^2)\|^{p/2+1} \\ &\qquad \cdot \|h'^{1/2}_{\alpha\beta}Du; L^2([0, t], L^2)\|^{p/2-1}\}, \end{aligned} \tag{3.57}_+$$

$$\begin{aligned} &\|\chi_-|u|^p; l^\infty(L^{q_1}([0, t], L^1))\| \\ &\quad \leqq C\{t^{1/q_1}\|\chi_- u; L^\infty([0, t], L^2)\|^p + \|\chi_- u; L^\infty([0, t], L^2)\|^{p/2+1} \\ &\qquad \cdot \|\chi_- Du; l^\infty(L^2([0, t], L^2))\|^{p/2-1}\}, \end{aligned} \tag{3.57}_-$$

for all $t \in [0, T)$, with $\alpha > 0$ and $1/q_1 = p/4 - \frac{1}{2}$ ($< \frac{1}{4}$).

(2) *Let V satisfy* (3.4) *with* $2 \leqq p < 3$. *Let* $T > 0$, *and let* $u_0 \in L^2$ *be such that* $(1+x_+)^{\beta/2} u_0 \in L^2$, *with* $\beta = \frac{1}{2} - 1/(2p)$. *Then* (1.1) *with initial data* $u(0) = u_0$ *has at most one solution satisfying* (3.5), (3.55), $(3.56)_\pm$, *and* $(3.17)_\pm$. *Any such solution satisfies* $(3.16)_\pm$ *with* $1/q_1 = p/4 - \frac{1}{2}$.

Proof. Part 1. It follows from $(3.56)_\pm$ that the last norm in $(3.57)_+$ is finite. The "+" part of the statement and $(3.57)_+$ follow immediately from Lemma 3.4 with $r = p$, $h = h_{\alpha\beta}$ and with $h_1 = c\chi_+(1+x)^{1/4}$ for $x \geqq 0$, the value of β being adjusted for that purpose. The "−" part and $(3.57)_-$ follow similarly from Lemma 3.3 and the Hölder inequality in time.

Part 2. Part 2 follows immediately from part 1 and Proposition 3.2. □

It follows from Proposition 3.4, from Proposition 3.6, part 1, and from Proposition 3.5, parts 1 and 2, the latter applied with $h = h_{\alpha\beta}$, that for V satisfying (3.4) with $2 \leqq p < 3$, the norms of smooth solutions of (1.1) corresponding to (3.55), $(3.56)_\pm$, $(3.17)_\pm$ and therefore also the norms corresponding to $(3.16)_\pm$ satisfy a priori estimates in terms of the initial data, and more precisely of $\|(1+x_+)^{\beta/2} u_0\|_2$ with $\beta = \frac{1}{2} - 1/(2p)$. That result indicates that for $2 \leqq p < 3$, (3.5), $(3.16)_\pm$, $(3.17)_\pm$ with $1/q_1 = p/4 - \frac{1}{2}$ and $\frac{4}{3} < q < 6$ define a suitable class for existence and uniqueness provided only $(1+x_+)^{\beta/2} u_0 \in L^2$. Similarly, (3.5), (3.55), $(3.56)_\pm$, $(3.17)_\pm$ with $1/q_1 = p/4 - \frac{1}{2}$ and $\frac{4}{3} < q < 6$ define another (smaller) suitable class.

The value of p in Proposition 3.6 is restricted to $p < 3$. This limitation comes from (actually expresses) the condition $q_1 > 4$ in $(3.16)_\pm$. Furthermore, the value of β that appears in the condition on u_0 increases with p. We will obtain better results in both respects by the second method, described below, of ensuring condition $(3.16)_\pm$.

PROPOSITION 3.7. *Let* $T > 0$ (*possibly* $T = \infty$). *Let* $p \geqq 2$.

(1) *Let u satisfy* (3.5) *and the conditions*

$$(1+x_+)^{1/8} u \in L^\infty_{\mathrm{loc}}([0, T), L^2), \tag{3.58}$$

$$\chi_+ u \in L^q_{\mathrm{loc}}([0, T), L^\infty), \tag{3.59}_+$$

$$\chi_- u \in l^\infty(L^q_{\mathrm{loc}}([0, T), L^\infty)), \tag{3.59}_-$$

with $q > \max(\frac{4}{3}, 4(p-2))$. *Then* $(3.16)_\pm$ *is satisfied with* $q_1 = q/(p-2) > 4$, *and for all* $t \in [0, T)$, *the following estimates hold*:

$$\begin{aligned} &\|\chi_+(1+x)^{1/4}|u|^p;\ L^{q_1}([0, T], L^1)\| \\ &\quad \leqq \|\chi_+(1+x)^{1/8} u;\ L^\infty([0, t], L^2)\|^2 \|\chi_+ u;\ L^q([0, t], L^\infty)\|^{p-2}, \end{aligned} \tag{3.60}_+$$

$$\begin{aligned} &\|\chi_-|u|^p;\ l^\infty(L^{q_1}([0, t], L^1))\| \\ &\quad \leqq \|\chi_- u;\ L^\infty([0, t], L^2)\|^2 \|\chi_- u;\ l^\infty(L^q([0, t], L^\infty))\|^{p-2}. \end{aligned} \tag{3.60}_-$$

Furthermore $(3.17)_\pm$ *hold with the same q and obvious estimates.*

(2) *Let V satisfy* (3.26) *with* $2 \leqq p < \frac{7}{2}$. *Let* $T > 0$, *let* $u_0 \in L^2$, *and let u be a solution of* (1.1) *with initial data* $u(0) = u_0$ *satisfying* (3.58), $(3.59)_\pm$ *with*

$$\max(\tfrac{4}{3}, 4(p-2)) < q < 6, \tag{3.61}$$

and in addition $Du \in l^\infty(L^2_{\mathrm{loc}}([0, T), L^2))$. *Then, for all* $t \in [0, T)$, *the norms of* $\chi_+ u$ *in* $L^q([0, t], L^\infty)$ *and of* $\chi_- u$ *in* $l^\infty(L^q([0, t], L^\infty))$ *are estimated in terms of t and of the norms of* $(1+x_+)^{1/8} u$ *in* $L^\infty([0, t], L^2)$, *and of* Du *in* $l^\infty(L^2, ([0, t], L^2))$.

(3) *Let V satisfy* (3.4) *with* $2 \leqq p < \frac{7}{2}$, *let* $T > 0$, *and let* $u_0 \in L^2$ *satisfy* $(1+x_+)^{1/8} u_0 \in L^2$. *Then* (1.1) *with initial data* $u(0) = u_0$ *has at most one solution satisfying* (3.5), (3.58),

and $(3.59)_{\pm}$ *with* q *satisfying* (3.61). *Any such solution satisfies* $(3.16)_{\pm}$ *with* $q_1 = q/(p-2)$ *and* $(3.17)_{\pm}$.

Proof. Part 1. Part 1 and $(3.60)_{\pm}$ follow immediately from the Hölder inequality.

Part 2. We prove the result in successive intervals $[s_j, s_{j+1}]$ covering the interval $[0, t]$, with $s_0 = 0$. In each such interval I, u satisfies the integral equation (3.29) with $s = s_j$. We proceed in two steps. The first step consists of estimating u in $l^\infty(L^q(I, L^\infty))$. For that purpose, we take a $\mathscr{C}^3$ function ψ with compact support, to be chosen more precisely later. Let χ be the characteristic function of the support of ψ. If u is a solution of (1.1), then ψu satisfies

$$\partial_t(\psi u) + D^3(\psi u) = 3D(\psi' Du) + \psi''' u + D(\psi V'(u)) - \psi' V'(u) \tag{3.62}$$

and therefore also the integral equation

$$\begin{aligned} \psi u(t) &= \chi U(t-s)\psi u(s) + \chi \int_s^t d\tau\, U(t-\tau) \\ &\quad \cdot \{3D(\psi' Du(\tau)) + \psi''' u(\tau) + D(\psi V'(u(\tau))) - \chi' V'(u(\tau))\} \\ &= \chi U(t-s)\psi u(s) + \sum_{1 \leqq i \leqq 4} J_i(t) \end{aligned} \tag{3.63}$$

where we have used the support properties of ψ and where the quantities $J_i(t)$, $i = 1, \cdots, 4$ are the contributions of the four terms in the last bracket.

We now estimate all terms in $L^q(I, L^\infty)$. The free term is estimated as before by (2.14) and the Hölder inequality as

$$\| U(\cdot - s)\psi u(s); L^q(I, L^\infty)\| \leqq C|I|^{1/q-1/6} \|u(s)\|_2. \tag{3.64}$$

To estimate J_1, we rewrite it as

$$J_1(t) = 3\chi e^{-x} \int_s^t d\tau \exp[-(t-\tau)(D-1)^3](D-1)e^x \psi' Du(\tau). \tag{3.65}$$

Furthermore,

$$(D-1)e^x \psi' Du = D(e^x \psi' Du - e^x \psi' u) + e^x(\psi' + \psi'')u. \tag{3.66}$$

The contributions to $J_1(t)$ of the first and second terms in the right-hand side of (3.66) are estimated by Lemma 2.2 (see especially (2.20) and (2.19), respectively) and the Hölder inequality as

$$\begin{aligned} \|J_1; L^q(I, L^\infty)\| &\leqq Ce^{|I|} \|\chi e^{-x}\|_\infty \{\|\psi'\, e^x\|_\infty (|I|^{1/q-1/6} \|\chi Du; L^2(I, L^2)\| \\ &\qquad + |I|^{1/q+1/3} \|\chi u; L^\infty(I, L^2)\|) \\ &\qquad + \|(\psi' + \psi'')\, e^x\|_\infty |I|^{1/q+5/6} \|\chi u; L^\infty(I, L^2)\|\} \\ &\leqq C(|I|)|I|^{1/q-1/6} (\|\chi Du; L^2(I, L^2)\| + \|\chi u; L^\infty(I, L^2)\|) \end{aligned} \tag{3.67}$$

where $C(|I|)$ depends on ψ in a translation-invariant manner. In particular, the exponentials in the various norms cancel up to a constant factor because of the support properties of ψ.

We estimate J_2 by (2.15) and the Hölder inequality as

$$\|J_2; L^q(I, L^\infty)\| \leqq C|I|^{1/q+5/6} \|\psi'''\|_\infty \|\chi u; L^\infty(I, L^2)\|. \tag{3.68}$$

We estimate J_3 and J_4 simply by using (2.10) and (2.9) so that

$$\|J_3(t) + J_4(t)\|_\infty \leqq C(|I|) \int_0^t d\tau\, |t-\tau|^{-3/4} \|\chi V'(u(\tau))\|_1 \tag{3.69}$$

where we have taken advantage of the support properties of χ and ψ', and $C(|I|)$ depends on ψ in a translation invariant manner. Taking the norm in $L^q(I)$ and using the Young and Hölder inequalities, we obtain

$$\|J_3+J_4;\ L^q(I,L^\infty)\| \leqq C(|I|)|I|^{1/4-(p-2)}\|\chi u;\ L^\infty(I,L^2)\|^2\|\chi u;\ L^q(I,L^\infty)\|^{p-1}. \tag{3.70}$$

Recalling the estimates (3.64), (3.67), (3.68), and (3.70), we obtain

$$\begin{aligned}\|\psi u;\ L^q(I,L^\infty)\| \leqq C(|I|)\{&|I|^{1/q-1/6}(\|\chi Du;\ L^2(I,L^2)\|+\|u;\ L^\infty(I,L^2)\|)\\ &+|I|^{1/4-(p-2)/q}\|u;\ L^\infty(I,L^2)\|^2\|\chi u;\ L^q(I,L^\infty)\|^{p-1}\}.\end{aligned} \tag{3.71}$$

We now choose a function $\psi_0\in\mathscr{C}^3$ with $0\leqq\psi_0\leqq 1$, $\psi_0(x)=1$ for $|x|\leqq\frac{1}{2}$ and $\psi_0(x)=0$ for $|x|\geqq 1$, and for all $j\in\mathbb{Z}$, we define ψ_j by $\psi_j(x)=\psi_0(x-j)$ so that for all $j\in\mathbb{Z}$,

$$\chi_j\leqq\psi_j\leqq\bar{\chi}_j\leqq\chi_{j-1}+\chi_j+\chi_{j+1}$$

where $\bar{\chi}_j$ is the characteristic function of the support $[j-1, j+1]$ of ψ_j. We now apply (3.71) with $\psi=\psi_j$ and take the supremum over j to obtain

$$\begin{aligned}\|u;\ l^\infty(L^q(I,L^\infty))\| \leqq C(|I|)\{&|I|^{1/q-1/6}(\|Du;\ l^\infty(L^2(I,L^2))\|+\|u;\ L^\infty(I,L^2)\|)\\ &+|I|^{1/4-(p-2)/q}\|u;\ L^\infty(I,L^2)\|^2\|u;\ l^\infty(L^q(I,L^\infty))\|^{p-1}\}.\end{aligned} \tag{3.72}$$

Now let

$$y(t)=\|u;\ l^\infty(L^q([s,t],L^\infty))\|. \tag{3.73}$$

It follows from (3.72) that $y(\cdot)$ is a continuous (actually Hölder continuous with exponent $\varepsilon=\min(1/q-\frac{1}{6},\frac{1}{4}-(p-2)/q)$) function of t and tends to zero when t decreases to s. Furthermore, (3.72) with $I=[s,t]$ yields an inequality of the form

$$y(t)\leqq a(t)+b(t)y(t)^{p-1} \tag{3.74}$$

where $a(t)$ and $b(t)$ tend to zero as $|t-s|^\varepsilon$ when t decreases to s. By an elementary and standard argument, this inequality implies that $y(t)$ is estimated a priori for $|t-s|$ sufficiently small, depending only on the norms of u in $L^\infty(\cdot,L^2)$ and of Du in $l^\infty(L^2(\cdot,L^2))$. Applying that argument in successive intervals $[s_j,s_{j+1}]$ yields the required a priori estimate of u in $l^\infty(L^q(\cdot,L^\infty))$.

The second step of the proof consists of estimating χ_+u in $L^q(\cdot,L^\infty)$, taking advantage of the estimate already obtained in the first step. For that purpose, we take $h=h_{\alpha 0}$ for some $\alpha>0$ and estimate $h^{1/2}u$ by inserting (2.28) of Lemma 2.3 into the integral equation (3.29) to obtain for $t\in I$

$$\begin{aligned}|h^{1/2}u(t,x)| \leqq{}& |h^{1/2}U(t-s)u(s)|+C(|I|)\int_s^t d\tau\,|t-\tau|^{-3/4}\\ &\cdot\Big\{\int_{y\geqq 0}dy\,(1+y)^{1/4}h^{1/2}|u(\tau,y)|^{p+1}\\ &\quad+\int_{y\leqq 0}dy\,\exp(-\gamma|x-y|/2)h^{1/2}|u(\tau,y)|^{p+1}\Big\}\\ \equiv{}& |h^{1/2}U(t-s)u(s)|+J_+(x)+J_-(x)\end{aligned} \tag{3.75}$$

(cf. (3.33)). By the same method as in the proof of Proposition 3.4, we estimate

$$\begin{aligned}&\|J_+;\ L^q(I,L^\infty)\|\\ &\quad\leqq C(|I|)|I|^{1/4-(p-2)/q}\|\chi_+(1+\cdot)^{1/8}u;\ L^\infty(I,L^2)\|^2\|\chi_+u;\ L^q(I,L^\infty)\|^{p-1},\end{aligned} \tag{3.76}$$

$$\|\chi_+ J_-;\, L^q(I, L^\infty)\| \leqq C(|I|)|I|^{1/4-(p-2)/q}\|\chi_- u;\, L^\infty(I, L^2)\|^2 \cdot \sum_{j\leqq 0} \exp\left(\frac{\gamma j}{2}\right)\|\chi_j\chi_- u;\, L^q(I, L^\infty)\|^{p-1}, \tag{3.77}$$

so that by collecting (3.64), (3.76), and (3.77)

$$\begin{aligned}\|\chi_+ u;\, L^q(I, L^\infty)\| &\leqq C|I|^{1/q-1/6}\|u;\, L^\infty(I, L^2)\| \\ &\quad + C(|I|)|I|^{1/4-(p-2)/q}\|(1+x_+)^{1/8}u;\, L^\infty(I, L^2)\| \\ &\quad \times\{\|\chi_+ u;\, L^q(I, L^\infty)\|^{p-1} + \|\chi_- u;\, l^\infty(L^q(I, L^\infty))\|^{p-1}\}.\end{aligned} \tag{3.78}$$

Therefore we again obtain an inequality of the type (3.74), for the quantity

$$y_+(t) = \|\chi_+ u;\, L^q([s, t], L^\infty)\|.$$

The end of the proof is then identical with that in the first step.

Part 3. This part follows immediately from part 1 and Proposition 3.2. □

It follows from Proposition 3.7, part 2 and from Proposition 3.5, part 1, and part 2 applied with $h = h_{\alpha 1/4}$, that for smooth solutions of (1.1) with V satisfying (3.4) and $2 \leqq p < \frac{7}{2}$, the norms corresponding to (3.58), $(3.59)_\pm$, and therefore, by Proposition 3.7, part 1, the norms corresponding to $(3.16)_\pm$ and $(3.17)_\pm$ with (3.61) and $q_1 = q/(p-2)$, satisfy a priori estimates in terms of the initial data, and more precisely in terms of $\|(1+x_+)^{1/8}u_0\|_2$. That result indicates that for $2 \leqq p < \frac{7}{2}$, (3.5), $(3.16)_\pm$, and $(3.17)_\pm$ with (3.61) and $q_1 = q/(p-2)$ define a suitable class for existence and uniqueness, provided only $(1+x_+)^{1/8}u_0 \in L^2$. Similarly (3.5), (3.58), $(3.59)_\pm$, (3.61) define another (smaller) suitable class. Note also that Proposition 3.7 improves over Proposition 3.6 as regards both the range of p and the value of β.

An interesting question left open at this stage is to determine sufficient smoothness conditions under which (I.1) and the a priori estimates of Proposition 3.5 hold true. Such conditions could be used, in particular, to exhibit uniqueness classes other than those defined in Propositions 3.1, 3.2, 3.6, and 3.7. Now the use of (I.1) requires that u satisfy (3.38). Although we are not able to derive (I.1) under that smoothness condition alone, we can go some way in that direction. For that purpose, we need a family of mollifiers $\varphi \in \mathscr{C}_0^\infty(\mathbb{R}, \mathbb{R}^+)$ with φ even and $\|\varphi\|_1 = 1$. We will eventually let φ tend to $\mathbb{1}$ in the sense that we will let μ tend to infinity in the one-parameter family $\{\varphi_\mu\}$ defined by $\varphi_\mu(x) = \mu\varphi_1(\mu x)$ for some fixed φ_1.

We can then prove the following result.

PROPOSITION 3.8. *Let V satisfy* (3.4) *with* $0 \leqq p \leqq 4$. *Let* $T > 0$, *let* $u_0 \in L^2$, *and let u be a solution of* (1.1) *with initial data* $u(0) = u_0$ *satisfying* (3.38). *Then*:

(1) *For any* $h \in \mathscr{C}^3$ *with compact support, for all* $t \in [0, T)$, *the following limit exists*:

$$\begin{aligned}&\exists \lim_{\varphi\to\mathbb{1}} 2\int_0^t d\tau\, \langle hDu_\varphi,\, V'(u_\varphi) - \varphi * V'(u)\rangle(\tau) \\ &= \langle u, hu\rangle(t) - \langle u_0, hu_0\rangle + 3\int_0^t d\tau\, \langle Du, h'Du\rangle(\tau) \\ &\quad - \int_0^t d\tau\, \langle u, h'''u\rangle(\tau) + 2\int_0^t d\tau \int h'(uV'(u) - V(u))(\tau)\end{aligned} \tag{3.79}$$

where φ is a mollifier, $u_\varphi = \varphi * u$, and all terms in the right-hand side are defined by absolutely convergent integrals.

(2) *If* $p \leqq 2$, *the identity* (I.1) *holds for all* $h \in \mathscr{C}^3$ *with compact support.*

(3) *Assume that* $p<4$ *and that the identity* (I.1) *holds for all* $h \in \mathscr{C}^3$ *with compact support. Then* $Du \in l^\infty(L^2_{\mathrm{loc}}([0, T), L^2))$. *Furthermore, for all* $\psi \in \mathscr{C}^2$ *with compact support, for all* $t \in [0, T)$, *the following limit exists*:

$$\exists \lim_{a \to \pm\infty} 3 \int_0^t d\tau \, \langle Du, \psi_a Du \rangle(\tau) = \|\psi\|_1 l_1^{\pm}(t) \tag{3.80}$$

where ψ_a *is defined by* $\psi_a(x) = \psi(x-a)$. *The functions* $l_1^{\pm}(t)$ *are nondecreasing in* t *and satisfy*

$$l_1^+(t) - l_1^-(t) = \|u(t)\|_2^2 - \|u_0\|_2^2 \tag{3.81}$$

for all $t \in [0, T)$. *The identity* (I.1) *holds for all* $h \in \mathscr{C}^3$ *with compactly supported* h' *if and only if* $l_1^{\pm}(t) = 0$ *for all* $t \in [0, T)$.

Proof. Part 1. We first remark that under the assumptions made here, $V'(u) \in L^1_{\mathrm{loc}}([0, T); L^2_{\mathrm{loc}})$. In fact, let χ be the characteristic function of a bounded interval. Then by Lemma 3.3,

$$\|\chi V'(u)\|_2 \leqq \|\chi |u|^{p+1}\|_2 \leqq C\{\|\chi u\|_2^{p+1} + \|\chi u\|_2^{1+p/2} \|\chi Du\|_2^{p/2}\}, \tag{3.82}$$

which belongs to $L^1_{\mathrm{loc}}([0, T))$ for $p \leqq 4$.

We now turn to the proof of (3.79). The function u_φ satisfies the equation

$$\partial_t u_\varphi + D^3 u_\varphi = D(\varphi * V'(u)). \tag{3.83}$$

Furthermore, $u_\varphi \in L^\infty_{\mathrm{loc}}([0, T), H^k)$ and $D(\varphi * V'(u)) \in L^1_{\mathrm{loc}}([0, T), H^k_{\mathrm{loc}})$ for any nonnegative integer k. By the same computation as in the proof of Lemma 2.5 (see (2.44)), we obtain for any $t \in [0, T)$

$$\begin{aligned}
&\langle u_\varphi, h u_\varphi \rangle(t) - \langle \varphi * u_0, h(\varphi * u_0) \rangle + 3 \int_0^t d\tau \, \langle Du_\varphi, h' Du_\varphi \rangle(\tau) - \int_0^t d\tau \, \langle u_\varphi, h''' u_\varphi \rangle(\tau) \\
&\quad = 2 \int_0^t d\tau \, \langle u_\varphi, hD(\varphi * V'(u)) \rangle \\
&\quad = 2 \int_0^t d\tau \, \langle u_\varphi, hDV'(u_\varphi) \rangle + 2 \int_0^t d\tau \, \langle u_\varphi, hD(\varphi * V'(u) - V'(u_\varphi)) \rangle \\
&\quad = -2 \int_0^t d\tau \int h'(u_\varphi V'(u_\varphi) - V(u_\varphi)) + 2 \int_0^t d\tau \, \langle h' u_\varphi + hDu_\varphi, (V'(u_\varphi) - \varphi * V'(u)) \rangle.
\end{aligned} \tag{3.84}$$

We then let φ tend to $\mathbb{1}$ wherever possible in (3.84), using the fact that the convolution with φ tends to $\mathbb{1}$ strongly in H^k for any $k \geqq 0$ and in L^r for any r, $1 \leqq r < \infty$. Under the assumptions made, all terms in the first member of (3.84) converge to the obvious limits. Next we consider the first term in the last member. From the identity

$$u_\varphi V'(u_\varphi) - V(u_\varphi) - (uV'(u) - V(u)) = (u_\varphi - u) \int_0^1 d\lambda \, \tilde{u} V''(\tilde{u}) \tag{3.85}$$

with $\tilde{u} = \lambda u_\varphi + (1-\lambda)u$, we obtain, with χ being the characteristic function of Support h' and $\bar{\chi}$ that of Support h' + Support φ,

$$\|\chi(\cdot)\|_1 \leqq C \|\chi(u_\varphi - u)\|_2 \|\bar{\chi} |u|^{p+1}\|_2, \tag{3.86}$$

which converges to zero in $L^1_{\mathrm{loc}}([0, T))$ by (3.82) and the Dominated Convergence Theorem.

The same argument shows that the contribution of $h'u_\varphi$ in the last integral in (3.84) tends to zero as φ tends to $\mathbb{1}$, so that the only term left uncontrolled at this stage in (3.84) is the contribution of hDu_φ to that same integral. This proves (3.79).

Part 2. In the same way as in part 1, we prove that $V'(u_\varphi)-\varphi * V'(u)$ now converges to zero in $L^2_{\text{loc}}([0, T), L^2_{\text{loc}})$. In fact,

$$V'(u_\varphi)-V'(u)=(u_\varphi-u)\int_0^1 d\lambda\ V''(\tilde{u})$$

so that by Lemma 3.3 (see (3.82)), with χ the characteristic function of a bounded interval I, and $\bar{\chi}$ that of $I+\text{Support }\varphi$,

$$\|\chi(V'(u_\varphi)-V'(u))\|_2 \leqq C\|u_\varphi-u\|_2\{\|\bar{\chi}u\|_2^p+\|\bar{\chi}u\|_2^{p/2}\|\bar{\chi}Du\|_2^{p/2}\},$$

which tends to zero in $L^2_{\text{loc}}([0, T))$ for $p\leqq 2$. Consequently, the integral in the left-hand side of (3.79) tends to zero for compactly supported h. This completes the proof of part 2.

Part 3. Let $\psi\in\mathscr{C}^2$ have compact support $[-R, R]$. Without loss of generality we can assume that $(0\leqq)\psi'^2\leqq C\psi$ and that $\|\psi\|_1=1$. We define $h_a^{\pm}$ by

$$h_a^{\pm}(x)=\pm\int_{\mp\infty}^{x} dy\,\psi_a(y)$$

so that $h_a^{\pm}(x)=1$ for $x\gtrless a\pm R$ and $h_a^{\pm}(x)=0$ for $x\lesseqgtr a\mp R$. We apply the identity (I.1) with $h=h_a^+h_b^-$ and $b-a\geqq 2R$ to obtain

$$\begin{aligned}&\langle u(t), hu(t)\rangle-\langle u_0, hu_0\rangle+3\int_0^t d\tau\,\langle Du, (\psi_a-\psi_b)Du\rangle(\tau)\\ &-\int_0^t d\tau\,\langle u, (\psi_a''-\psi_b'')u\rangle(\tau)\\ &=-2\int_0^t d\tau\int(\psi_a-\psi_b)(uV'(u)-V(u))(\tau).\end{aligned}\tag{3.87}$$

By the same computation as in the proof of Lemma 3.2, applied with h replaced by ψ and $W(u)=uV'(u)-V(u)$ (see especially (3.43)) and by using the support properties of ψ, we can estimate the contribution of ψ_b to the right-hand side of (3.87) as

$$\begin{aligned}&\left|2\int_0^t d\tau\int\psi_b(uV'(u)-V(u))(\tau)\right|\\ &\quad\leqq C\int_0^t d\tau\,\{\|\chi_b u\|_2^{p+2}+\|\chi_b u\|_2^{2+p/2}\langle Du, \psi_b Du\rangle^{p/4}\}(\tau)\\ &\quad\leqq C\int_0^t d\tau\,\|\chi_b u\|_2^{p+2}+Ct^{1-p/4}\left\{\int_0^t d\tau\,\langle Du, \psi_b Du\rangle(\tau)\right\}^{p/4}\\ &\qquad\times\|\chi_b u; L^\infty([0, t], L^2)\|^{2+p/2}\end{aligned}\tag{3.88}$$

where χ_b is the characteristic function of $[b-R, b+R]$. From (3.87), (3.88), it follows that for fixed a, the integral

$$y(t)=\int_0^t d\tau\,\langle Du, \psi_b Du\rangle(\tau)$$

satisfies an inequality of the type

$$y(t) \leqq A(t) + B(t) y(t)^{p/4}$$

with $A(t)$ and $B(t)$ uniformly bounded with respect to b and therefore, for $p < 4$, $y(t)$ is uniformly bounded with respect to b. By the same argument with a and b interchanged, this proves that $Du \in l^\infty(L^2_{\mathrm{loc}}([0, T), L^2))$. We next let b tend to infinity for fixed a in (3.87). From the previous fact, from (3.88), from the fact that $\|\chi_b u\|_2$ tends to zero pointwise in t as $b \to \infty$, and from the Dominated Convergence Theorem, it follows that the contribution of ψ_b'' to the left-hand side of (3.87) and that of ψ_b to the right-hand side tend to zero as $b \to \infty$. We then obtain the existence of the limit:

$$\begin{aligned} \exists \lim_{b\to\infty} 3 \int_0^t d\tau \, \langle Du, \psi_b Du\rangle(\tau) &= \langle u, h_a^+ u\rangle(t) - \langle u_0, h_a^+ u_0\rangle \\ &\quad + 3 \int_0^t d\tau \, \langle Du, \psi_a Du\rangle(\tau) - \int_0^t d\tau \, \langle u, \psi_a'' u\rangle(\tau) \\ &\quad + 2 \int_0^t d\tau \int \psi_a (uV'(u) - V(u))(\tau). \end{aligned} \tag{3.89}$$

The same argument applies to the limit $a \to -\infty$ for fixed b. The fact that the limit in (3.80) depends on ψ only through $\|\psi\|_1$ is easily seen by defining $h_a^\pm$ with two different functions ψ_a^+, ψ_b^- at a and b, the only constraint being that $\|\psi^+\|_1 = \|\psi^-\|_1$ in order that h have compact support. The right-hand side of (3.89) then depends only on ψ_a^+ and is therefore independent of the specific choice of ψ^- satisfying that constraint.

Relation (3.81) follows by taking successively the limits $b \to +\infty$ and $a \to -\infty$ in (3.87), or $a \to -\infty$ in (3.89).

The last statement follows in one direction by applying the identity (I.1) for the (compactly supported) function $h_a^+ h_b^-$ with h_a^+, h_b^- defined as before, and letting $b \to \infty$ and $a \to -\infty$, and in the opposite direction by applying (I.1) with $h = h_a^\pm$ and letting a tend to $\pm\infty$. □

Remark 3.5. In the proof of part 3, we could have replaced the assumption that V satisfies (3.4) with $p < 4$ by the weaker semiboundedness condition (3.46) of Proposition 3.5, at the expense of using again the estimates of Lemma 3.2. We have refrained from doing so because we are interested in part 3 of Proposition 3.8 only in the more restricted context considered here.

Remark 3.6. Relation (3.81) suggests that under the weak assumptions made on the smoothness of solutions, the local H^1 norm may act as a source at $+\infty$ and/or a sink at $-\infty$ for the L^2 norm of the solutions, and some kind of boundary condition at $\pm\infty$ must be imposed to exclude that effect. Note also that if we are interested only in the estimates of the kind provided by Proposition 3.5, only the condition on $l_1^+(t)$ is important, since then for the relevant h, (I.1) is replaced by an inequality in the right direction.

Remark 3.7. If $u \in (L^\infty_{\mathrm{loc}} \cap \mathscr{C}_W)([0, T), H^1)$, then (I.1) holds for any $h \in \mathscr{C}^3$ with compactly supported h' and condition (3.4) with $p \leqq 4$ is not needed for that purpose. (See Proposition 4.1 below.)

Using Propositions 3.5 and possibly 3.8, we finally exhibit different uniqueness classes for (1.1) than those described in Proposition 3.1 for $p \leqq 2$ and in Propositions 3.2, 3.6, and 3.7 for $p \geqq 2$. We consider first the case $p \leqq 2$.

PROPOSITION 3.9. *Let V satisfy* (3.4) *with* $0<p\leqq 2$, *let* $(1+x_+)^{\beta/2}u_0\in L^2$ *with* $\beta=1/p-\frac{1}{4}$, *and let* $T>0$ (*possibly* $T=\infty$). *Then* (1.1) *with initial data* $u(0)=u_0$ *has at most one solution satisfying* (3.38), *that is, condition*

$$\liminf_{j\to+\infty}\|\chi_j Du;\ L^2([0,t],L^2)\|=0 \tag{3.90}$$

for all $t\in T$, *and condition* (3.7) *with* (3.9), (3.10). *Any such solution satisfies*

$$(1+x_+)^{\beta/2}u\in L^\infty_{\mathrm{loc}}([0,T),L^2).$$

Proof. The result follows immediately from Propositions 3.1, 3.5, and 3.8 and from Remark 3.6. □

We next turn to the case $p\geqq 2$.

PROPOSITION 3.10. *Let V satisfy* (3.4) *with* $p\geqq 2$, *let* $u_0\in L^2$, *and let* $T>0$. *Then* (1.1) *with initial data* $u(0)=u_0$ *has at most one solution u satisfying* (3.38), (I.1) *for all* $h\in\mathscr{C}^3$ *with compactly supported* h', *and either of the two sets of conditions*:

(1) $2\leqq p<3$, $(1+x_+)^{\beta/2}u_0\in L^2$ *for* $\beta=\frac{1}{2}-1/(2p)$, *and u satisfies* $(3.17)_\pm$.

(2) $2\leqq p<\frac{7}{2}$, $(1+x_+)^{1/8}u_0\in L^2$, *and u satisfies* $(3.59)_\pm$ *with q satisfying* (3.61).

Proof. The result follows immediately from Propositions 3.2, 3.5, and 3.6 under conditions 1, and from Propositions 3.2, 3.5, and 3.7 under conditions 2. □

Remark 3.8. We had to keep (3.90) in Proposition 3.9 and (I.1) in Proposition 3.10 as independent assumptions, since we are unable to derive (I.1) for solutions of (1.1) satisfying only (3.38). An unpleasant consequence is that it is unclear whether the uniqueness classes defined in Propositions 3.9 and 3.10 are suitable for existence proofs. In fact, all the relevant norms (including that of Du in $l^\infty(L^2(\cdot,L^2))$) are estimated a priori in terms of the initial data for smooth solutions, but conditions such as (3.90) are not preserved when taking weak-star limits.

4. Uniqueness of H^1 solutions. In this section, we derive our results on the uniqueness of H^1 solutions of the GKdV equation (1.1), more precisely of solutions in $(L^\infty_{\mathrm{loc}}\cap C_W)([0,T),H^1)$ for some $T>0$. Such solutions can also be called finite energy solutions: for such a solution, the energy

$$E(u)=\tfrac{1}{2}\|Du\|_2^2+\int dx\ V(u) \tag{4.1}$$

is well defined for each $t\in[0,T)$ and formally conserved, i.e., $E(u(t))=\text{const}$. Conversely, energy conservation and a semiboundedness property of V imply an a priori estimate of the solution in $L^\infty([0,T),H^1)$ (see Proposition 4.4 below for more precise statements). H^1 solutions satisfy Propositions 3.5 and 3.8 with stronger conclusions and weaker assumptions on V. We state that and a related result immediately for later reference.

PROPOSITION 4.1. *Let* $u_0\in H^1$, *let* $T>0$, *and let* $u\in(L^\infty_{\mathrm{loc}}\cap C_W)([0,T),H^1)$ *be a solution of* (1.1) *with initial data* $u(0)=u_0$. *Then*:

(1) *The identity* (I.1) *holds for all* $h\in\mathscr{C}^3$ *with compactly supported* h'. *The L^2-norm of u is conserved, namely,* $\|u(t)\|_2=\|u_0\|_2$ *for all* $t\in[0,T)$.

(2) *Let* $h\in\mathscr{C}^3(\mathbb{R},\mathbb{R}^+)$ *satisfy* $0\leqq h'\leqq c_1(1+h)$, $h'''\leqq c_3(1+h)$ *and let* $h^{1/2}u_0\in L^2$. *Then* $h^{1/2}u\in L^\infty_{\mathrm{loc}}([0,T),L^2)$, $h'^{1/2}Du\in L^2_{\mathrm{loc}}([0,T),L^2)$ *and for all* $t\in[0,T)$, $h^{1/2}u$ *is estimated in* $L^\infty([0,t],L^2)$ *and* $h'^{1/2}Du$ *is estimated in* $L^2([0,t],L^2)$ *in terms of t, of* $\|(1+h)^{1/2}u_0\|_2$ *and of* $\|Du;\ L^\infty([0,t),L^2)\|$. *In addition* (I.1) *holds for that specific h.*

(3) $u\in l^2(L^\infty_{\mathrm{loc}}([0,T),L^2))$ *and for all* $t\in[0,T)$, u *is estimated in* $l^2(L^\infty([0,t],L^2))$ *in terms of t and of* $\|u;\ L^\infty([0,t],H^1)\|$.

Proof. Part 1. By a Sobolev inequality and elementary arguments, $u \in L^\infty_{\mathrm{loc}}([0, T), L^\infty)$ and the map $u \to V'(u)$ is continuous from H^1 to L^2. Part 1 is then proved in the same way as Proposition 3.8. In particular, the left-hand side of (3.79) tends to zero when φ tends to $\mathbb{1}$.

Part 2. Part 2 is proved in the same way as Proposition 3.5, with the last term in (I.1) now simply estimated as

$$\left|2\int h'(uV'(u)-V(u))\right| \leqq M(\|u\|_\infty)\langle u, h'u\rangle \tag{4.2}$$

for some locally bounded function M. Note also that the statement on $h'^{1/2}Du$ adds new information only if h' is unbounded.

Part 3. Let $\psi \in \mathscr{C}^3$ have compact support, and let χ be the characteristic function of the support of ψ. From (I.1) with $h = \psi$ we obtain by direct estimation

$$\begin{aligned}\langle u, \psi u\rangle(t) \leqq \langle u_0, \psi u_0\rangle + 3c_1 \int_0^t d\tau\, \langle Du, \chi Du\rangle(\tau) \\ + \int_0^t d\tau\, \|\chi u(\tau)\|_2^2 (c_3 + 2c_1 M(\|u(\tau)\|_\infty))\end{aligned} \tag{4.3}$$

where $c_1 = \|\psi'\|_\infty$, $c_3 = \|\psi'''\|_\infty$, and $M(\rho) = \sup_{|\rho'| \leqq \rho} |V''(\rho')|$. We now choose a function $\psi_0 \in \mathscr{C}^3$ with $0 \leqq \psi_0 \leqq 1$, $\psi_0(x) = 1$ for $|x| \leqq \frac{1}{2}$ and $\psi_0(x) = 0$ for $|x| \geqq 1$, and for all $j \in \mathbb{Z}$, we define ψ_j by $\psi_j(x) = \psi_0(x - j)$. We apply (4.3) with $\psi = \psi_j$, take the norms in $L^\infty([0, t])$, and take the sum over j to obtain

$$\begin{aligned}\|u; l^2(L^\infty([0, t], L^2))\|^2 &\leqq 2\|u_0\|_2^2 + 6c_1 \int_0^t d\tau\, \|Du(\tau)\|_2^2 \\ &\quad + 2(c_3 + 2c_1 M(\|u; L^\infty([0, t], L^\infty)\|)) \int_0^t d\tau\, \|u(\tau)\|_2^2 \\ &= 2\|u\|_2^2\{1 + t(c_3 + 2c_1 M(\|u; L^\infty([0, t], L^\infty)\|))\} \\ &\quad + 6c_1 \|Du; L^2([0, t], L^2)\|_2^2\end{aligned} \tag{4.4}$$

where the last equality follows from the conservation of the L^2-norm. Part 3 then follows from (4.4). □

We now turn to our basic uniqueness result. As in § 3, it is based on a linear inequality for a suitable norm of the difference of two solutions. That inequality now arises from a variant of (I.1). In this section we will often assume (in addition to the general assumptions $V \in \mathscr{C}^2$, $V(0) = V'(0) = 0$) that V'' is absolutely continuous ($\equiv AC$) with $V''(0) = 0$ and with locally bounded (Radon-Nikodym or distributional) derivative V'''. We will, however, refrain from assuming $V \in \mathscr{C}^3$, since we do not want to exclude such cases as $V''(u) = |u|$.

PROPOSITION 4.2. *Let V'' be AC with $V''(0) = 0$, and let V''' be locally bounded. Let $u_0 \in H^1$, and let $T > 0$ (possibly $T = \infty$). Then* (1.1) *with initial data $u(0) = u_0$ has at most one solution such that*

$$u \in (L^\infty_{\mathrm{loc}} \cap C_w)([0, T), H^1), \tag{4.5}$$

$$\chi_+ Du \in L^1_{\mathrm{loc}}([0, T), L^\infty). \tag{4.6}$$

Proof. Let u_1 and u_2 be two solutions satisfying (4.5), (4.6) with common initial data $u_1(0) = u_2(0) = u_0$, let $w = u_1 - u_2$, and let $h = h_{\alpha 0}$ for some $\alpha > 0$. The proof is based on a variant of (I.1) satisfied by w. For clarity, we give only the algebraic part

of the proof of that identity in differential form. Under the smoothness assumption (4.5), the functional analytic details can be easily filled in by the mollifier method of Propositions 3.8 and 4.1. From (3.1) and the same computation as in the proof of Lemma 2.5 (see (2.44)), we obtain

$$\partial_t\langle w, hw\rangle + 3\langle Dw, h'Dw\rangle - \langle w, h'''w\rangle = 2\langle w, hD(\tilde{V}''w)\rangle. \tag{4.7}$$

For $x \geqq 0$, we use the identity

$$2wD(\tilde{V}''w) = D(w^2\tilde{V}'') + w^2(D\tilde{V}'')$$

to obtain

$$2\int_0^\infty dx\, hwD(\tilde{V}''w) = \langle w, h\chi_+(D\tilde{V}'')w\rangle - \langle w, h'\chi_+\tilde{V}''w\rangle - (hw^2\tilde{V}'')(0). \tag{4.8}$$

For $x \leqq 0$, we use the identity

$$2wD(\tilde{V}''w) = 2D(w^2\tilde{V}'') - 2w(Dw)\tilde{V}''$$

to obtain

$$2\int_{-\infty}^0 dx\, hwD(\tilde{V}''w) = -2\langle Dw, h\chi_-\tilde{V}''w\rangle - 2\langle w, h'\chi_-\tilde{V}''w\rangle + 2(hw^2\tilde{V}'')(0). \tag{4.9}$$

Adding (4.8) and (4.9), we obtain

$$\begin{aligned} 2\langle w, hD(\tilde{V}''w)\rangle = {} & \langle w, h\chi_+(D\tilde{V})w\rangle - \langle w, h'\chi_+\tilde{V}''w\rangle - 2\langle w, h'\chi_-\tilde{V}''w\rangle \\ & - 2\langle Dw, h\chi_-\tilde{V}''w\rangle + \tilde{V}''(0)(\langle w, h'\chi_-w\rangle + 2\langle Dw, h\chi_-w\rangle) \end{aligned} \tag{4.10}$$

(the argument denoted as zero in (4.8)-(4.10) is the space variable x). Using the properties of h, namely, $\chi_-h' = \alpha\chi_-h$ and $0 \leqq \chi_+h' \leqq \alpha\chi_+h$, and an elementary quadratic inequality, from (4.10) we obtain

$$\begin{aligned} 2\langle w, hD(\tilde{V}''w)\rangle \leqq {} & (\|\chi_+D\tilde{V}''\|_\infty + 3\alpha\|\tilde{V}''\|_\infty + 2\alpha^{-1}\|\chi_-\tilde{V}''\|_\infty^2) \\ & \times\langle w, hw\rangle + 2\langle Dw, h'\chi_-Dw\rangle. \end{aligned} \tag{4.11}$$

The result now follows from (4.7) and (4.11) by Gronwall's inequality, insofar as $\tilde{V}'' \in L^2_{\mathrm{loc}}([0, T), L^\infty)$ and $\chi_+D\tilde{V}'' \in L^1_{\mathrm{loc}}([0, T), L^\infty)$. The former property follows from (4.5), actually with L^∞ instead of L^2 in time, while the latter follows from (4.6) and the additional assumption on V''. □

Remark 4.1. Condition (4.6) of Proposition 4.2 may seem inadequate for a later treatment of the existence problem along the lines sketched in the Introduction, since L^1 is not the dual of a Banach space. However, we can replace L^1 by L^q with any $q \geqq 1$ in (4.6) and the subsequent a priori estimates will actually hold with $q = 6$.

The remaining part of this section is devoted to the analysis of conditions (4.6) for H^1 solutions. As will become clear in the subsequent discussion, the fact that no difficulty arises from the lack of local regularity, namely, for high values of p in (3.4), is in contrast to the case of L^2 solutions considered in the previous section. On the other hand, there remains a difficulty, caused by the lack of decrease at infinity in space, giving rise in particular to the lower bound $p > \frac{3}{2}$ if no further assumption is made on u_0 than $u_0 \in H^1$. An important tool in overcoming that difficulty is a second

well-known identity [10], satisfied by the solutions of (1.1), which we will henceforth refer to as (I.2) and which we formulate in the following integral form:

$$\langle Du, hDu\rangle(t)+2\int hV(u(t))+3\int_0^t d\tau\,\langle D^2u, h'D^2u\rangle(\tau)+\int_0^t d\tau\int h'V'(u(\tau))^2$$

$$=\langle Du_0, hDu_0\rangle+2\int hV(u_0) \tag{I.2}$$

$$+\int_0^t d\tau\left\{\langle Du, h'''Du\rangle(\tau)+2\int h'''V(u(\tau))-4\langle Du, h'DV'(u)\rangle(\tau)\right\}.$$

Clearly (I.2) makes sense for $h\in\mathscr{C}^3$ with compactly supported h' provided

$$u\in(L^\infty_{\mathrm{loc}}\cap C_w)([0,T), H^1)\cap L^2_{\mathrm{loc}}([0,T), H^2_{\mathrm{loc}}). \tag{4.12}$$

Identity (I.2) can be used to derive weighted H^1 estimates for solutions of (1.1) in the same way as (I.1) was used to derive weighted L^2 estimates for those solutions. We could therefore proceed immediately to the derivation of those estimates, in close analogy with Proposition 3.5, by assuming (I.2), and leave its derivation for a later stage, in analogy with our treatment of (I.1) in § 3. However, the situation for (I.2) is better than that for (I.1), and we are almost able to derive (I.2) for solutions of (1.1) satisfying (4.12). We will therefore reverse the order and begin with the derivation of (I.2).

PROPOSITION 4.3. *Let $T>0$, let $u_0\in H^1$, and let u be a solution of* (1.1) *with initial data $u(0)=u_0$ satisfying* (4.12). *Then*:

(1) *For any $h\in\mathscr{C}^3$ with compact support, for any $t\in[0,T)$, u satisfies* (I.2). *Furthermore, $u\in l^\infty(L^2_{\mathrm{loc}}([0,T), H^2))$ and for all $\psi\in\mathscr{C}^2$ with compact support, for all $t\in[0,T)$, the following limits exist*:

$$\exists\lim_{a\to\pm\infty} 3\int_0^t d\tau\,\langle D^2u, \psi_aD^2u\rangle(\tau)=2\|\psi\|_1 l_2^{\pm}(t) \tag{4.13}$$

where ψ_a is defined by $\psi_a(x)=\psi(x-a)$. The functions $l_2^{\pm}(t)$ are nondecreasing in t and satisfy

$$l_2^+(t)-l_2^-(t)=E(u(t))-E(u_0) \tag{4.14}$$

for all $t\in[0,T)$. Identity (I.2) *holds for all $h\in\mathscr{C}^3$ with compactly supported h' if and only if $l_2^{\pm}(t)=0$ for all $t\in[0,T)$.*

(2) *Assume in addition either that*

$$\liminf_{j\to\pm\infty}\|\chi_jD^2u;\, L^2([0,t], L^2)\|=0 \tag{4.15}$$

for all $t\in[0,T)$, or that V satisfies (3.4) *for some $p\geqq 2$ and for all $\rho\in[-1,1]$. Then* (I.2) *holds for all $h\in\mathscr{C}^3$ with compactly supported h'.*

Proof. Part 1. We first remark that the map $u\to V'(u)$ is continuous from H^1 to H^1. In fact, since $DV'(u)=V''(u)Du$, it suffices to prove that the map $u\to V''(u)$ is continuous from H^1 to L^∞. This follows from the continuous embedding $H^1\subset L^\infty$ and from the fact that V'' is uniformly continuous on bounded intervals.

We now turn to the proof of (I.2). We use the same mollifier method as in the proof of Proposition 3.6. From (3.83), we obtain by the same computation as in the

proof of Lemma 2.5 (see (2.44))

$$\begin{aligned}&\langle Du_\varphi, hDu_\varphi\rangle(t)-\langle \varphi * Du_0, h(\varphi * Du_0)\rangle+3\int_0^t d\tau\,\langle D^2u_\varphi, h'D^2u_\varphi\rangle(\tau)\\&\quad-\int_0^t d\tau\,\langle Du_\varphi, h'''Du_\varphi\rangle(\tau)=2\int_0^t d\tau\,\langle Du_\varphi, hD^2(\varphi * V'(u))\rangle(\tau).\end{aligned}\tag{4.16}$$

The integrand on the right-hand side of (4.16) can be rewritten as

$$\begin{aligned}\langle Du_\varphi, hD^2(\varphi * V'(u))\rangle&=\langle D^2u_\varphi, hD(V'(u_\varphi)-\varphi * V'(u))\rangle\\&\quad-\langle Du_\varphi, h'D(\varphi * V'(u))\rangle\\&\quad+\langle D^2u_\varphi, h'V'(u_\varphi)\rangle+D^3u_\varphi, hV'(u_\varphi)\rangle.\end{aligned}\tag{4.17}$$

Again using (3.83) to re-express the last term, we obtain by an elementary computation

$$\begin{aligned}\cdots&=-\partial_t\int hV(u_\varphi)-\tfrac{1}{2}\langle\varphi * V'(u), h'(\varphi * V'(u))\rangle+\int h'''V(u_\varphi)\\&\quad-\langle Du_\varphi, h'D(\varphi * V'(u)+V'(u_\varphi))\rangle\\&\quad+\langle D(\varphi * V'(u)), h(V'(u_\varphi)-\varphi * V'(u))\rangle\\&\quad+\langle D^2u_\varphi, hD(V'(u_\varphi)-\varphi * V'(u))\rangle.\end{aligned}\tag{4.18}$$

We substitute (4.18) into the right-hand side of (4.16) and take the limit $\varphi\to\mathbb{1}$ wherever possible. By using only the assumptions that $u\in L^\infty_{\mathrm{loc}}([0,T), H^1)$ and $h\in\mathscr{C}^3$ with compactly supported h', we obtain from the fact that u_φ tends to u, and that $\varphi * V'(u)$ and $V'(u_\varphi)$ tend to $V'(u)$ in H^1, that all terms tend to the obvious limits. Possible exceptions are the first integral on the left-hand side of (4.16) and the contribution of the last term in (4.18). If in addition $u\in L^2_{\mathrm{loc}}([0,T), H^2_{\mathrm{loc}})$, the former integral also tends to the obvious limit, so that

$$\begin{aligned}&\exists\lim_{\varphi\to\mathbb{1}} 2\int_0^t d\tau\,\langle D^2u_\varphi, hD(V'(u_\varphi)-\varphi * V'(u))\rangle(\tau)\\&\quad=\langle Du, hDu\rangle(t)+2\int hV(u(t))+3\int_0^t d\tau\,\langle D^2u, h'D^2u\rangle(\tau)\\&\qquad+\int_0^t d\tau\int h'V'(u(\tau))^2-\langle Du_0, hDu_0\rangle-2\int hV(u_0)\\&\qquad-\int_0^t d\tau\left\{\langle Du, h'''Du\rangle(\tau)+2\int h'''V(u(\tau))-4\langle Du, h'DV'(u)\rangle(\tau)\right\}.\end{aligned}\tag{4.19}$$

Furthermore, the integral on the left-hand side of (4.19) tends to zero when $\varphi\to\mathbb{1}$ if in addition h has compact support, thereby proving (I.2) in that special case. The remaining statements in part 1 of the proposition are proved by the same method as part 3 of Proposition 3.8, using, in particular, (I.2) with $h=h_a^+h_b^-$ and letting a (respectively, b) vary and/or tend to $-\infty$ (respectively, $+\infty$) for fixed b (respectively, a). We omit the details.

Part 2. Under assumption (4.15), we obtain from (4.13) that $l_2^\pm(t)=0$ and (I.2) for $h\in\mathscr{C}^3$ with compactly supported h' follow from the last statement of part 1. To prove (I.2) under the additional assumption on V we start from (4.19) and prove directly that the left-hand side is zero. Since $D^2u\in l^\infty(L^2([0,t], L^2))$ by part 1, it suffices

to prove that $D(V'(u_\varphi)-\varphi * V'(u))$ tends to zero in $l^1(L^2([0,t],L^2))$. Since $V'(u_\varphi)$ and $\varphi * V'(u)$ tend to $V'(u)$ in H^1 for each t, that fact will follow from the Dominated Convergence Theorem and from an estimate of $DV'(u)$ and $DV'(u_\varphi)$ uniformly in φ. We concentrate on $DV'(u)$. The same estimate will hold for $DV'(u_\varphi)$. By the additional assumption on V, it is sufficient to estimate $|u|^p Du$. Now from Lemma 3.3, we obtain

$$\|\chi_j|u|^p Du\|_2 \leqq C(\|\chi_j Du\|_2\|\chi_j u\|_2^p + \|\chi_j Du\|_2^{1+p/2}\|\chi_j u\|_2^{p/2}) \tag{4.20}$$

so that by the Hölder inequality

$$\begin{aligned}&\||u|^p Du;\, l^1(L^2([0,t],L^2))\| \\ &\leqq C\{\|Du;\, L^2([0,t],L^2)\|\,\|u;\, l^{2p}(L^\infty([0,t],L^2))\| \\ &\quad + \|Du;\, l^{p+2}(L^{p+2}([0,t],L^2))\|^{1+p/2}\|u;\, l^p(L^\infty([0,t],L^2))\|^{p/2}\}.\end{aligned} \tag{4.21}$$

The result now follows from the fact that $Du \in L^\infty([0,t],L^2) \subset l^q(L^q([0,t],L^2))$ for all q, $2 \leqq q \leqq \infty$, and that $u \in l^2(L^\infty([0,t],L^2)) \subset l^q(L^\infty([0,t],L^2))$ for all $q \geqq 2$ by Proposition 4.1, part 3. □

We now turn to the derivation of weighted $L^\infty(H^1)$ estimates for solutions of (1.1), in close analogy with Proposition 3.5.

PROPOSITION 4.4. *Let V satisfy the conditions*

$$\lim_{|\rho|\to\infty} |\rho|^{-6} V_-(\rho) = 0, \tag{4.22}$$

$$V_-(\rho) \leqq C|\rho|^{p+2} \quad \text{for all } \rho \in [-1,1], \tag{4.23}$$

for some p, $0 \leqq p < 4$, and some $C \geqq 0$. Let $T > 0$, let $u_0 \in H^1$, and let u be a solution of (1.1) *with initial data $u(0) = u_0$, satisfying* (4.12), *and such that* (I.2) *holds for all $h \in \mathscr{C}^3$ with compactly supported h'. Then*:

(1) *The energy is conserved, namely, $E(u(t)) = E(u_0)$ for all $t \in T$, and u is estimated in $L^\infty([0,T),H^1)$ in terms of $\|u_0;\, H^1\|$.*

(2) *$D^2u \in l^\infty(L^2_{\mathrm{loc}}([0,T),L^2))$, and for all $t \in [0,T)$, D^2u is estimated in $l^\infty(L^2([0,t],L^2))$ in terms of t and $\|u_0;\, H^1\|$.*

(3) *Let $h \in \mathscr{C}^3(\mathbb{R},\mathbb{R}^+)$ satisfy $0 \leqq h' \leqq c_1(1+h)$ and $|h'''| \leqq c_3(1+h)$. Let $h^{1/2}Du_0 \in L^2$, $hV_+(u_0) \in L^1$, and let $h^{\nu/2}u$, $(1+h)^{-1/2}h'u \in L^\infty_{\mathrm{loc}}([0,T),L^2)$, where $\nu = (4-p)/(4+p)$. Then $h^{1/2}Du \in L^\infty_{\mathrm{loc}}([0,T),L^2)$, $hV_+(u) \in L^\infty_{\mathrm{loc}}([0,T),L^1)$, $h'^{1/2}D^2u \in L^2_{\mathrm{loc}}([0,T);L^2)$, $h'^{1/2}V'(u) \in L^2_{\mathrm{loc}}([0,T);L^2)$, and for all $t \in [0,T)$, the corresponding norms in the interval $[0,t]$ are estimated in terms of t, of $\|(1+h)^{1/2}Du_0\|_2$, of $\|hV_+(u_0)\|_1$, and of the norms of $(1+h)^{\nu/2}u$ and $(1+h)^{-1/2}h'u$ in $L^\infty([0,t],L^2)$. In addition,* (I.2) *holds for that specific choice of h.*

Proof. Part 1. Part 1 follows from (I.2) with $h \equiv 1$ and from Lemma 3.2 applied with $h \equiv 1$ and $W = V_-$.

Next we consider an h with compactly supported h'. From (I.2) with h replaced by $(1+h)$ we obtain

$$\begin{aligned}&\langle Du, (1+h)Du\rangle(t) + 2\int (1+h)V(u(t)) + 3\int_0^t d\tau\,\langle D^2u, h'D^2u\rangle(\tau) \\ &\quad + \int_0^t d\tau \int h'V'(u(\tau))^2 \leqq \langle Du_0, (1+h)Du_0\rangle + 2\int (1+h)V(u_0) \\ &\quad + \int_0^t d\tau \Big\{(c_3 + 4c_1\|V''(u);\, L^\infty([0,t],L^\infty)\|)\langle Du, (1+h)Du\rangle(\tau) \\ &\qquad + 2c_3\int (1+h)|V(u(\tau))|\Big\}.\end{aligned} \tag{4.24}$$

From Lemma 3.2 (see especially (3.41)) applied with $W = V_-$ and h replaced by $(1+h)$, it follows that the quantity

$$F(u) = \langle Du, (1+h)Du\rangle + \int (1+h)V(u) + C \tag{4.25}$$

is positive in the interval $[0, t]$ for some constant $C \geqq 0$ depending only on the norms of $(1+h)^{\nu/2}u$ and $(1+h)^{-1/2}h'u$ in $L^\infty([0, t], L^2)$. Furthermore, F satisfies an inequality of the type

$$\begin{aligned} &F(u(t)) + 3\int_0^t d\tau\, \langle D^2u, h'D^2u\rangle(\tau) + \int_0^t d\tau \int h'V'(u(\tau))^2 \\ &\leqq F(u_0) + A + B\int_0^t d\tau\, F(u(\tau)) \end{aligned} \tag{4.26}$$

by Lemma 3.2 again and by part 1, with A and B depending only on $\|u_0; H^1\|$ and on the norms of $(1+h)^{\nu/2}u$ and $(1+h)^{-1/2}h'$ in $L^\infty([0, t], L^2)$.

Part 2. Part 2 now follows from (4.26) applied with the same h as in the proof of Proposition 3.5, part 1. (Note, in particular, that for that choice, A and B in (4.26) and C in (4.25) are estimated in terms of $\|u_0; H^1\|$ and that $F(u(t))$ is estimated in terms of $\|u_0; H^1\|$ uniformly in t.)

Part 3. Part 3 follows from (4.25) and (4.26) by approximating h by a sequence of functions h_N with compactly supported h'_N as in the proof of Proposition 3.5, part 2, estimating the corresponding F_N uniformly in N by (4.26) and the Gronwall inequality, taking the limit $N \to \infty$, and again applying Lemma 3.2. We omit the details. □

Remark 4.2. In the applications, the fact that $h^{\nu/2}u$ and $(1+h)^{-1/2}h'u$ belong to $L^\infty_{\text{loc}}([0, T), L^2)$ will follow from the obvious assumptions on u_0 through Proposition 4.1. We have kept these properties as assumptions in Proposition 4.4 to avoid additional assumptions on h that would be irrelevant for the main estimate (4.24).

We now come back to the analysis of condition (4.6). We first give an elementary treatment thereof that is inspired by the estimates of § 3 and makes it possible to exhibit a different uniqueness class.

PROPOSITION 4.5. *Let V satisfy* (3.4) *for some $p \geqq 1$ and all $\rho \in [-1, 1]$. Let $T > 0$, and let $u_0 \in H^1$ be such that $(1+x_+)^{\beta/2}u_0 \in L^2$ and $(1+x_+)^{\gamma/2}Du_0 \in L^2$ for some $\beta, \gamma \geqq 0$ satisfying*

$$(p+1)(\beta+\gamma) \geqq 1. \tag{4.27}$$

(1) *Let u be a solution of* (1.1) *with initial data $u(0) = u_0$, satisfying* (4.5) *and in addition*

$$(1+x_+)^{\gamma/2}Du \in L^\infty_{\text{loc}}([0, T), L^2). \tag{4.28}$$

Then u also satisfies

$$\chi_+ Du \in L^6_{\text{loc}}([0, T), L^\infty) \tag{4.29}$$

and for all $t \in [0, T)$, χ_+Du is estimated in $L^6([0, t], L^\infty)$ in terms of t, of $\|(1+x_+)^{\beta/2}u_0\|_2$, and of the norm of $(1+x_+)^{\gamma/2}Du$ in $L^\infty([0, t], L^2)$.

(2) *In addition let V'' be AC with $V''(0)=0$ and V''' locally bounded. Then* (1.1) *with initial data $u(0)=u_0$ has at most one solution satisfying* (4.5) *and* (4.28). *Any such solution satisfies* (4.29).

Proof. Part 1. Let u be a solution of (1.1) with $u(0)=u_0$, satisfying (4.5) and (4.28). Du satisfies

$$\begin{aligned} Du(t) &= U(t)Du_0+\int_0^t d\tau\, U(t-\tau)D(V''(u)Du)(\tau) \\ &\equiv U(t)Du_0+J(t). \end{aligned} \tag{4.30}$$

We estimate the free term by (2.14) as

$$\|U(\cdot)Du_0;\, L^6(\mathbb{R}^+,\, L^\infty)\| \leqq C\|Du_0\|_2. \tag{4.31}$$

Using the fact that $\chi_+ \leqq h_{\alpha 0} \leqq C$ and the estimate (2.35) we obtain for $\chi_+ J$

$$\|\chi_+ J;\, L^\infty([0,\, t],\, L^\infty)\| \leqq C(t)t^{1/4}\|(1+x_+)^{1/4}V''(u)Du;\, L^\infty([0,\, t],\, L^1)\|.$$

By (3.4) for $|\rho| \leqq 1$, the last norm is estimated as

$$M(\|u;\, L^\infty([0,\, t],\, L^\infty)\|)\|(1+x_+)^{1/4}|u|^p Du;\, L^\infty([0,\, t],\, L^1)\|.$$

Next we estimate for each t

$$\|(1+x_+)^{1/4}|u|^p Du\|_1 \leqq \|(1+x_+)^{\beta/2}u\|_2\|(1+x_+)^{\gamma/2}Du\|_2\|(1+x_+)^{\varepsilon}u^2\|_\infty^{(p-1)/2}$$

with $(p-1)\varepsilon = \frac{1}{2}-\beta-\gamma$. For $\varepsilon \leqq 0$, the last norm is estimated by $(2\|u\|_2\|Du\|_2)^{1/2}$. For $\varepsilon > 0$, we obtain by derivation and integration

$$\|(1+x_+)^{\varepsilon}u^2\|_\infty \leqq C(\|(1+x_+)^{\beta/2}u\|_2^2+\|(1+x_+)^{\beta/2}u\|_2\|(1+x_+)^{\gamma/2}Du\|_2)$$

provided $\varepsilon - 1 \leqq \beta$ and $\varepsilon \leqq (\beta+\gamma)/2$. The latter condition is identical to (4.27) while the former reduces to

$$(p+1)(\beta+\gamma)+(p-1)(\beta-\gamma+2) \geqq 1$$

and follows from (4.27) provided $\gamma \leqq 2$, a condition that we can impose without loss of generality. Collecting the previous estimates, we obtain

$$\begin{aligned} &\|\chi_+ J;\, L^\infty([0,\, t],\, L^\infty)\| \\ &\quad \leqq C(t)t^{1/4}M(\|u;\, L^\infty([0,\, t],\, L^\infty)\|) \\ &\qquad \times\{\|(1+x_+)^{\beta/2}u;\, L^\infty([0,\, t],\, L^2)\|^p\|(1+x_+)^{\gamma/2}Du;\, L^\infty([0,\, t],\, L^2)\| \\ &\qquad\quad +\|(1+x_+)^{\beta/2}u;\, L^\infty([0,\, t],\, L^2)\|^{(p+1)/2} \\ &\qquad\qquad \times\|(1+x_+)^{\gamma/2}Du;\, L^\infty([0,\, t],\, L^2)\|^{(p+1)/2}\}. \end{aligned} \tag{4.32}$$

The last two norms in (4.32) are controlled, respectively, by Proposition 4.1, part 2 applied with $h=h_{0\beta}$, and by (4.28).

Part 1 then follows from (4.31) and (4.32).

Part 2. Part 2 follows immediately from part 1 and Proposition 4.2. □

Proposition 4.5 enables us to cover the case of the ordinary KdV equation corresponding to $p=1$. In that case, (4.27) reduces to $\beta+\gamma \geqq \frac{1}{2}$. On the other hand, it does not yield uniqueness under the assumption that $u_0 \in H^1$ only, even for higher values of p. We now turn to a more elaborate method for ensuring (4.6), which will remedy that defect. The general result stated in Proposition 4.6 below contains as a special case a result obtained previously by Tsutsumi [22], according to which uniqueness holds for H^1 data provided $p > \frac{3}{2}$ in (3.4). Since that result has special interest and since its proof is simpler than that of the more general one, we state it separately.

PROPOSITION 4.6. *Let V'' be AC with $V''(0)=0$ and V''' locally bounded, and let V satisfy* (3.4) *for some $p\geqq 1$ and all $\rho\in[-1,1]$. Let $T>0$ and let $u_0\in H^1$. Then*:

(1) *Let $p>\frac{3}{2}$. Let u be a solution of* (1.1) *with initial data $u(0)=u_0$, satisfying* (4.12). *Then u also satisfies*

$$Du\in L^6_{\mathrm{loc}}([0,T),L^\infty) \tag{4.33}$$

and for all $t\in[0,T)$, Du is estimated in $L^6([0,t],L^\infty)$ in terms of t, of the norms of u in $L^\infty([0,t],H^1)$ and of D^2u in $l^\infty(L^2([0,t],L^2))$.

(2) *Let $p>\frac{3}{2}$. Then* (1.1) *with initial data $u(0)=u_0$ has at most one solution satisfying* (4.12). *Any such solution satisfies* (4.33).

In addition let u_0 satisfy $(1+x_+)^{\beta/2}u_0\in L^2$ and $(1+x_+)^{\gamma/2}Du_0\in L^2$ for some β, $\gamma\geqq 0$ satisfying

$$(3-p)_+\beta+(p-1)(2+3\min(\beta,\gamma))>1. \tag{4.34}$$

Then:

(3) *Let u be a solution of* (1.1) *with initial data $u(0)=u_0$, satisfying* (4.12) *and* (4.28). *Then u also satisfies* (4.29), *and for all $t\in[0,T)$, χ_+Du is estimated in $L^6([0,t],L^\infty)$ in terms of t, of $\|(1+x_+)^{\beta/2}u_0\|_2$ and of the norms of $(1+x_+)^{\gamma/2}Du$ in $L^\infty([0,t],L^2)$ and of D^2u in $l^\infty(L^2([0,t],L^2))$.*

(4) *Equation* (1.1) *with initial data $u(0)=u_0$ has at most one solution satisfying* (4.12) *and* (4.28). *Any such solution satisfies* (4.29).

Proof. Part 1. Let u be a solution of (1.1) with $u(0)=u_0$ satisfying (4.12). By Proposition 4.1, part 3, $u\in l^2(L^\infty_{\mathrm{loc}}([0,T),L^2))$. We estimate Du by using (4.30) and estimating the free term by (4.31). We then split J as follows:

$$J(t)=J_1(t)+J_2(t)=\int_0^t d\tau\, U(t-\tau)\{V'''(u)(Du)^2+V''(u)D^2u\}(\tau) \tag{4.35}$$

where J_1 and J_2 are the contributions of the terms in the last bracket. We estimate J_1 by the use of (2.9) as

$$\begin{aligned}\|J_1;L^\infty([0,t],L^\infty)\|&\leqq Ct^{2/3}\|V'''(u)(Du)^2;L^\infty([0,t],L^1)\|\\&\leqq Ct^{2/3}M_1(\|u;L^\infty([0,t],L^\infty)\|)\|Du;L^\infty([0,t],L^2)\|^2\end{aligned} \tag{4.36}$$

where $M_1(\rho)=\sup_{|\rho'|\leqq\rho}|V'''(\rho')|$. The last member of (4.36) is estimated in terms of the norm of u in $L^\infty([0,t],H^1)$. We now turn to J_2. Using the estimate (2.8), we can estimate J_2 as follows:

$$\begin{aligned}|J_2(t,x)|&\leqq\int_0^t d\tau\,|t-\tau|^{-1/3}\int dy\,A(x-y)(V''(u)D^2u)(\tau,y)\\&\quad+C\int_0^t d\tau\,|t-\tau|^{-1/4}\int_{y\geqq x}dy(1+y-x)^{-1/4}(V''(u)D^2u)(\tau,y)\\&=J_2^-(t,x)+J_2^+(t,x)\end{aligned} \tag{4.37}$$

where A is a bounded function with exponential decrease. Using the Hölder inequality in spaces $l^r(L^r(\cdot,L^r))$, we estimate J_2^- as

$$\begin{aligned}\|J_2^-;L^\infty([0,t],L^\infty)\|&\leqq Ct^{1/6}\|A;l^1(L^2)\|\,\|V''(u);L^\infty([0,t],L^\infty)\|\\&\quad\times\|D^2u;l^\infty(L^2([0,t],L^2))\|.\end{aligned} \tag{4.38}$$

The norm of $V''(u)$ is estimated in terms of that of u in the same space, and therefore of the norm of u in $L^\infty([0,t],H^1)$. The last norm in (4.38) is finite by Proposition 4.3,

part 1. This takes care of J_2^-. We now turn to the (more delicate) estimate of J_2^+. The function $|t|^{-1/4}(1+|x|)^{-1/4}$ belongs to $l^r(L^q(I, L^\infty))$ for any bounded interval I and $q<4<r$. Using that fact together with $1/q+1/r=\frac{1}{2}$, $r>4$, by the Hölder inequality we obtain

$$\begin{aligned}&\|J_2^+; L^\infty([0,t], L^\infty)\| \\ &\leqq Ct^{1/4-1/r}\|V''(u)D^2u; l^{\bar r}(L^{\bar q}([0,t], L^1))\| \\ &\leqq Ct^{1/4-1/r}\|V''(u); l^{\bar r}(L^r([0,t], L^2))\|\,\|D^2u; l^\infty(L^2([0,t], L^2))\|.\end{aligned} \tag{4.39}$$

The last norm is finite by Proposition 4.3, part 1, while the norm of V'' is estimated by

$$\|\cdot\| \leqq M(\|u; L^\infty([0,t], L^\infty)\|)\,\||u|^p; l^{\bar r}(L^r([0,t], L^2))\| \tag{4.40}$$

by the use of (3.4) for $|\rho|\leqq 1$. Now by Proposition 4.1, part 3, $u\in l^m(L^\infty(\cdot, L^2))$ for all m, $2\leqq m\leqq\infty$. Furthermore, it follows from the fact that $u\in L^\infty(\cdot, H^1)$ that $u\in l^s(L^s(\cdot, H^1))$ for all s, $2\leqq s\leqq\infty$, with

$$\|u; l^s(L^s([0,t], H^1))\| \leqq t^{1/s}\|u; L^\infty([0,t], H^1)\|. \tag{4.41}$$

We use Lemma 3.3, the Hölder inequality in time and in the discrete l^r variable and the previous remarks to estimate the last norm in (4.40) as

$$\begin{aligned}&\||u|^p; l^{\bar r}(L^r([0,t], L^2))\| \\ &\leqq C\|u; l^m(L^\infty([0,t], L^2))\|^{(p+1)/2}\|u; l^s(L^s([0,t], H^1))\|^{(p-1)/2}t^{1/r-(p-1)/(2s)} \\ &\leqq C\|u; l^2(L^\infty([0,t], L^2))\|^{(p+1)/2}\|u; L^\infty([0,t], H^1)\|^{(p-1)/2}t^{1/r}\end{aligned} \tag{4.42}$$

provided we can choose $m\geqq 2$, $s\geqq 2$ such that

$$\frac{p-1}{s}\leqq\frac{2}{r}, \tag{4.43}$$

$$\frac{p-1}{s}+\frac{p+1}{m}=\frac{2}{\bar r}. \tag{4.44}$$

It is convenient to choose m and s by imposing the additional condition

$$\frac{p+1}{m}-\frac{p-1}{s}=1$$

so that $(p+1)/m=\frac{3}{2}-1/r$, $(p-1)/s=\frac{1}{2}-1/r$. In that case, conditions $m\geqq 2$, $s\geqq 2$ both become equivalent to $p\geqq 2/\bar r$, which can be ensured with $r>4$ provided $p>\frac{3}{2}$ while condition (4.43) reduces to $r\leqq 6$ that can be imposed separately. The required estimate of J_2^+ now follows from (4.39), (4.40), and (4.42), from Proposition 4.1, part 3 and Proposition 4.3, part 1. Part 1 of this proposition follows from that estimate and from (4.31), (4.36), and (4.38).

Part 2. Part 2 follows immediately from part 1 and from Proposition 4.2.

Part 3. Because of part 1, we can assume that $p\leqq 3$ without loss of generality. We again estimate χ_+Du by using (4.30), and estimate the free term J_1 and J_2^- as in the proof of part 1. To estimate $\chi_+J_2^+$, we start from (4.37), decompose the contribution of the y integral as the sum of the contributions of unit intervals, and apply Lemma 3.3 in each interval, thereby obtaining for each $k\geqq 0$

$$\begin{aligned}\|\chi_kJ_2^+(t)\|_\infty \leqq M\int_0^t d\tau\,|t-\tau|^{-1/4}\sum_{j\geqq k}(1+j-k)^{-1/4}\|\chi_jD^2u(\tau)\|_2 \\ \times(\|\chi_ju(\tau)\|_2^p+\|\chi_ju(\tau)\|_2^{(p+1)/2}\|\chi_jDu(\tau)\|_2^{(p-1)/2})\end{aligned} \tag{4.45}$$

where M depends on $\|u; L^\infty([0,t], L^\infty)\|$ according to (4.40). We now factorize and estimate the integrand in (4.45) by using the fact that $(1+j-k)\leqq 1+j$ since $0\leqq k\leqq j$,

and we continue (4.45):

$$
\begin{aligned}
\cdots \leqq M \int_0^t d\tau \sum_{j \geqq k} & \{\|\chi_j D^2 u(\tau)\|_2\}\{|t-\tau|^{-1/4}\|\chi_j(1+x)^{\beta/2}u\|\}^{(3-p)/2} \\
& \times\{|t-\tau|^{-1/4}(\|\chi_j(1+x)^{\beta/2}u\|_2+\|\chi_j(1+x)^{\gamma/2}Du\|_2)\}^{(p-1)/2} \\
& \times\{(1+j-k)^{-\mu/2}\|\chi_j(1+x)^{\delta/2}u\|_2^{p-1}\}
\end{aligned} \tag{4.46}
$$

with $\delta = \min(\beta, \gamma)$ and

$$
2\mu = (3-p)\beta + 3(p-1)\delta + 1. \tag{4.47}
$$

We estimate the last member of (4.46) by using the Hölder inequality in time and in the discrete variable j, with the four successive brackets taken in $l^\infty(L^2)$, $l^2(L^2)$ $(\equiv L^2(l^2))$, $l^2(L^2)$, and $l^2(L^\infty)$, respectively, thereby continuing (4.46):

$$
\begin{aligned}
\cdots \leqq M & \|\chi_+ D^2 u;\, l^\infty(L^2([0,t],L^2))\|\, t^{1/4}\|\chi_+(1+x)^{\beta/2}u;\, L^\infty([0,t],L^2)\|^{(3-p)/2} \\
& \times(\|\chi_+(1+x)^{\beta/2}u;\, L^\infty([0,t],L^2)\| \\
& \quad +\|\chi_+(1+x)^{\gamma/2}Du;\, L^\infty([0,t],L^2)\|)^{(p-1)/2} \\
& \times\left\{\sum_{j\geqq k}(1+j-k)^{-\mu}\|\chi_j(1+x)^{\delta/2}u;\, L^\infty([0,t],L^2)\|^{2(p-1)}\right\}^{1/2}.
\end{aligned} \tag{4.48}
$$

The last factor in (4.48) is then estimated by

$$
\|\chi_+(1+x)^{\delta/2}u;\, l^2(L^\infty([0,t],L^2))\|^{p-1} \tag{4.49}
$$

either in an obvious way if $p \geqq 2$ since μ is positive, or by the Hölder inequality in the discrete variable provided

$$
\mu + p - 1 > 1,
$$

which coincides with (4.34). Now the norms appearing in (4.48), except for the last one, are controlled by Proposition 4.3, part 1, by Proposition 4.1, part 2 applied with $h = h_{0\beta}$ and by (4.28), while the norm in (4.49) is controlled in terms of the previous one by a straightforward extension of Proposition 4.1, part 3. The required estimate of $\chi_+ J_2^+$ now follows from (4.45), (4.46), (4.48), and (4.49). Part (3) of this proposition follows from that estimate and from (4.31), (4.36), and (4.38).

Part 4. Part 4 follows immediately from part 3 and Proposition 4.2. □

Remark 4.3. All the solutions of (1.1) considered in Proposition 4.5 and in parts (3) and (4) of Proposition 4.6 satisfy automatically the condition $(1+x_+)^{\beta/2}u \in L^\infty_{\mathrm{loc}}([0,T),L^2)$ by Proposition 4.1, part 2. Furthermore, it follows from the proofs of Propositions 4.5 and 4.6 that (4.29) and (4.33) can be supplemented with the statements that

$$
\chi_+ Du - \chi_+ U(\cdot)Du_0 \in L^\infty_{\mathrm{loc}}([0,T),L^\infty), \tag{4.50}
$$

$$
Du - U(\cdot)Du_0 \in L^\infty_{\mathrm{loc}}([0,T),L^\infty) \tag{4.51}
$$

with the same estimates as those given in connection with (4.29), (4.33).

Using Proposition 4.4, we can obtain sufficient conditions for the various norms appearing in the uniqueness classes, namely, the norms associated with (4.6), (4.12), and (4.28) to be a priori estimated in terms of the initial data, and we can exhibit a new uniqueness class.

PROPOSITION 4.7. *Let V satisfy the condition* (4.22) *and the condition* (3.4) *for some* p, $1 \leqq p < 4$, *and for all* $\rho \in [-1,1]$. *Let* $T > 0$, *and let* $u_0 \in H^1$ *satisfy* $(1+x_+)^{\beta/2}u_0 \in L^2$ *and* $(1+x_+)^{\gamma/2}Du_0 \in L^2$ *for some* $\beta, \gamma \geqq 0$ *satisfying* $\beta \geqq \nu\gamma$, *where* $\nu = (4-p)/(4+p)$, *and* $\gamma \leqq \beta + 2$.

(1) *Let u be a solution of* (1.1) *with initial data $u(0)=u_0$, satisfying* (4.12) *and in addition* (4.15) *if $p<2$. Then u satisfies $D^2u\in l^\infty(L^2_{\mathrm{loc}}([0,T),L^2))$ and* (4.28), *and for all $t\in[0,T)$, u is estimated in $L^\infty([0,t],H^1)$, D^2u is estimated in $l^\infty(L^2([0,t],L^2))$ and $(1+x_+)^{\gamma/2}Du$ is estimated in $L^\infty([0,t],L^2)$ in terms of t of $\|(1+x_+)^{\beta/2}u_0\|_2$ and of $\|(1+x_+)^{\gamma/2}Du_0\|_2$.*

(2) *In addition let V'' be AC with $V''(0)=0$ and V''' locally bounded, and let p, β, γ satisfy either* (4.27) *or* (4.34). *Then the equation* (1.1) *with initial data $u(0)=u_0$ has at most one solution satisfying* (4.12) *and in addition* (4.15) *if $p<2$. Any such solution satisfies* (4.28) *and* (4.29).

Proof. Part 1. Part 1 follows from Proposition 4.4. Assumption (3.4) implies (4.23) with the same p by integration. The statements and estimates on u in $L^\infty(\cdot,H^1)$ and D^2u in $l^\infty(L^2(\cdot,L^2))$ are a repetition of parts 1 and 2 of Proposition 4.4. The statements and estimates on $(1+x_+)^{\gamma/2}Du$ in $L^\infty(\cdot,L^2)$ follow from part 3 of Proposition 4.4 applied with $h=h_{\alpha\gamma}$ for some $\alpha>0$. The additional assumptions required on $h^{\nu/2}u$ and $(1+h)^{1/2}h'u$ follow from Proposition 4.1, part 2, applied with $h_{\alpha\beta}$ for some $\alpha>0$, under the conditions $\beta\geqq\nu\gamma$ and $\gamma\leqq\beta+2$. Finally, the assumption $hV_+(u_0)\in L^1$ follows from Lemma 3.2 applied with $W=V$.

Part 2. Part 2 follows from part 1 and from Proposition 4.5, part 2 and Proposition 4.6, parts 2 and 4. □

Proposition 4.7 indicates that under the additional assumptions (4.23), $\beta\geqq\nu\gamma$ and $\gamma\geqq\beta+2$, the uniqueness classes defined in Proposition 4.5, part 2 and Proposition 4.6, parts 2 and 4 are actually suitable for existence and uniqueness provided only $(1+x_+)^{\beta/2}u_0\in L^2$ and $(1+x_+)^{\gamma/2}Du_0\in L^2$, with the appropriate assumptions on V, β, and γ made in those propositions. That need not be the case, however, for the uniqueness class defined in Proposition 4.7 for $p<2$, since the condition (4.15) is not preserved when taking weak-star limits (cf. Remark 3.8).

We conclude this section with some comments on the assumptions of Propositions 4.5–4.7. In view of parts 1 and 2 of Proposition 4.6, parts (3) and (4) of that proposition and to a large extent Proposition 4.5 are interesting only for $1\leqq p\leqq\frac{3}{2}$, and, in particular, for the ordinary KdV equation corresponding to $p=1$. In that case, the interaction $V''(u)$ does not decrease sufficiently fast as x tends to $+\infty$ for H^1 data, and an additional decrease must be assumed on the data, either on u or on Du. Which of Propositions 4.5 and 4.6 is better depends on the values of β and γ. For $\gamma=0$, condition (4.34) is always weaker than (4.27), except in the limiting case $p=1$, $\beta=\frac{1}{2}$, which is allowed by (4.27) but not by (4.34). On the other hand, for low values of p and β, (4.27) yields better results through a better use of γ. For instance, for $p=1$, (4.27) together with $\beta\geqq\nu\gamma$ allows for $\beta=\frac{3}{16}$, $\gamma=\frac{5}{16}$, which is not covered by (4.34). This may be due in part to the fact that in Proposition 4.6, parts 3 and 4, we have not used the property $\chi_+(1+x_+)^{(\gamma-1)/2}D^2u\in L^2_{\mathrm{loc}}([0,T),L^2)$, which follows from Proposition 4.4, part 3, but we will not elaborate on that point.

Acknowledgments. We are grateful to Professor J. C. Saut for enlightening discussions. The second author is grateful to Professor K. Chadan for his kind hospitality at the Laboratoire de Physique Théorique et Hautes Energies at Orsay.

REFERENCES

[1a] J. L. BONA AND J. C. SAUT, *Singularités dispersives de solutions d'équations de type Korteweg-de Vries*, C. R. Acad. Sci. Paris, 303 (1986), pp. 101-103.

[1b] ———, *Dispersive blow up solutions of nonlinear dispersive wave equations*, in preparation.

[2] J. L. BONA AND L. R. SCOTT, *Solutions of the Korteweg-de Vries equation in fractional order Sobolev spaces*, Duke Math. J., 43 (1976), pp. 87-99.
[3] J. L. BONA AND R. SMITH, *The initial value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 555-601.
[4] A. COHEN MURRAY, *Solutions of the Korteweg-de Vries equation from irregular data*, Duke Math. J., 45 (1978), pp. 149-181.
[5] A. COHEN AND T. KAPPELER, *Solutions to the Korteweg-de Vries equation with initial profile in* $L^1_1(\mathbb{R}) \cap L^1_N(\mathbb{R}^+)$, SIAM J. Math. Anal., 18 (1987), pp. 991-1025.
[6] P. CONSTANTIN AND J. C. SAUT, *Local smoothing properties of dispersive equations*, Trans. Amer. Math. Soc., to appear.
[7] J. GINIBRE AND G. VELO, *Scattering theory in the energy space for a class of non linear Schrödinger equations*, J. Math. Pures Appl., 64 (1985), pp. 363-401.
[8] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, Berlin, New York, 1983.
[9] T. KATO, *On the Korteweg-de Vries equation*, Manuscripta Math., 28 (1979), pp. 89-99.
[10] ———, *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, in Studies in Applied Mathematics, V. Guillemin, ed., Advanced Mathematics Supplementary Studies, 18, Academic Press, New York, 1983, pp. 93-128.
[11] ———, *On nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987), pp. 113-129.
[12] S. N. KRUZHKOV AND A. V. FAMINSKII, *Generalized solutions of the Cauchy problem for the Korteweg-de Vries equation*, Math. USSR Sb., 48 (1984), pp. 391-421.
[13] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod Gauthier-Villars, Paris, 1969.
[14] R. M. MIURA, *The Korteweg-de Vries equation: a survey of results*, SIAM Rev., 18 (1976), pp. 412-459.
[15] T. MUKASA AND R. LINO, *On the global solutions for the simplest generalized Korteweg-de Vries equation*, Math. Japoni., 14 (1969), pp. 75-83.
[16] J. C. SAUT AND R. TEMAM, *Remarks on the Korteweg-de Vries equation*, Israel J. Math., 24 (1976), pp. 78-87.
[17] S. TANAKA, *Korteweg-de Vries equation: construction of solutions in terms of scattering data*, Osaka J. Math., 11 (1974), pp. 49-59.
[18] R. TEMAM, *Sur un problème non linéaire*, J. Math. Pures Appl., 48 (1969), pp. 159-172.
[19] M. TSUTSUMI, *On global solutions of the generalized Korteweg-de Vries equation*, Publ. Res. Inst. Math. Sci., 7 (1972), pp. 329-344.
[20] M. TSUTSUMI, T. MUKASA, AND R. LINO, *On the generalized Korteweg-de Vries equations*, Proc. Japan Acad. Ser. A Math. Sci., 46 (1970), pp. 921-925.
[21] M. TSUTSUMI AND T. MUKASA, *Parabolic regularization of the generalized Korteweg-de Vries equation*, Funkcial. Ekvac., 14 (1971), pp. 89-110.
[22] Y. TSUTSUMI, *On uniqueness of solutions in the energy space for the modified* KdV *equation*, unpublished manuscript.
[23] K. YAJIMA, *Existence of solutions of Schrödinger evolution equations*, Comm. Math. Phys., 110 (1987), pp. 415-426.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1426-1435, November 1989

008

ON A KIND OF PREDATOR-PREY SYSTEM*

DING SUNHONG†

Abstract. In this paper a kind of predator-prey system given in [*SIAM J. Appl. Math.*, 35 (1978), pp. 617-625] is considered. Utilizing the theory of ordinary differential equations, two theorems for a general predator-prey system are proved, completing the investigation of the predator-prey system.

Key words. predator-prey system, limit cycle, Poincaré-Bendixon theorem

AMS(MOS) subject classification. 34C

A predator-prey system is a simple and typical mathematical model of the ecological system. Recently, many ecologists and mathematicians have been paying attention to establishing many mathematical models for the predator-prey system, have been investigating these mathematical models, and have been interpreting these ecological systems from a scientific viewpoint [2], [5], [7], [8]. Unfortunately, it is not easy to investigate the predator-prey system completely. In this paper we give two theorems on the existence of unique periodic solutions and the nonexistence of periodic solutions for a kind of predator-prey system. Applying these two theorems, we complete the investigation of the predator-prey system, that has been done in [3], in the global parameter space.

The mathematical model of the predator-prey system given in [3] is

$$\begin{aligned} \dot{X} &= \gamma X(1-X/K) - YX^n/(a+X^n), \\ \dot{Y} &= Y(\mu X^n/(a+X^n) - D), \qquad n=1,2 \end{aligned} \tag{1}$$

where X and Y are functions of t, $\dot{X}=dX/dt$, $\dot{Y}=dY/dt$, and γ, a, μ, K, and D are positive parameters.

We consider a general form of (1):

$$\dot{X}=\varphi(X)-g(Y)\psi(X), \qquad \dot{Y}=h(Y)(\mu\psi(X)-D) \tag{2}$$

where

(H1) $\varphi(X)\in C^1(0,+\infty)$, $\varphi(0)=0$,
$\psi(X)\in C^1(0,+\infty)$, $\psi(0)=0$ and $\psi'(X)>0$;

(H2) $h(Y)\in C(0,+\infty)$, $h(0)=0$ and $h'(Y)>0$,
$g(Y)\in C(0,+\infty)$, $g(0)=0$ and $g'(Y)>0$;

(H3) μ and D are positive parameters.

Let $F(X)=\varphi(X)/\psi(X)$. In the domain

$$\Omega=\{(X,Y)\,|\,X>0,\ Y>0\},$$

(2) is equivalent to

$$\begin{aligned} \dot{X} &= \varphi(X)(F(X)-g(Y)) = G(X,Y), \\ \dot{Y} &= h(Y)(\mu\psi(X)-D) = H(X,Y). \end{aligned} \tag{3}$$

* Received by the editors November 11, 1987; accepted for publication (in revised form) June 9, 1988.
† Institute of Mathematical Sciences, Chengdu Branch, Academia Sinica, Chengdu, People's Republic of China.

According to the ecological meaning, we investigate (1)-(3) in the domain Ω only. Since $\psi'(X)>0$ as $X>0$, the equation $\mu\psi(X)-D=0$ has at most one positive solution λ; hence (2) or (3) has at most one singular point $P(\lambda, g^{-1}(F(\lambda)))$ in the domain Ω. Theorem 1 is on the existence of a unique periodic solution for (2) or (3).

THEOREM 1. *Let $P(\lambda, g^{-1}(F(\lambda)))$ be a unique singular point of* (3) *in the domain Ω, let conditions* (H1)-(H3) *be satisfied, let $F'(\lambda)\neq 0$, and let $F'(X)\psi(X)/(\mu\psi(X)-D)$ be nonincreasing in the intervals $(0,\lambda)$ and $(\lambda,+\infty)$. Then* (3) *has at most one stable limit cycle in the domain Ω.*

Proof. Define a function

$$V(X, Y)=\int_{\lambda}^{X}\frac{\mu\psi(\xi)-D}{\psi(\xi)}\,d\xi+\int_{g^{-1}(F(\lambda))}^{Y}\frac{g(\eta)-F(\lambda)}{h(\eta)}\,d\eta.$$

Then

$$\left(\frac{dV}{dt}\right)_{(3)}=(F(X)-F(\lambda))(\mu\psi(X)-D) \tag{4}$$

where $(dV/dt)_{(3)}$ denotes the total derivative of the function $V(X, Y)$ along a path of (3) corresponding to a solution $(X(t), Y(t))$. Since $F'(X)\psi(X)/(\mu\psi(X)-D)$ is nonincreasing in the intervals $(0,\lambda)$ and $(\lambda,+\infty)$ and $F'(\lambda)\neq 0$, then $F'(\lambda)>0$. On the contrary, assume $F'(\lambda)<0$; then $\lim_{X\to\lambda_{+0}}F'(X)\psi(X)/(\mu\psi(X)-D)=-\infty$. Because $F'(X)\psi(X)/(\mu\psi(X)-D)$ is nonincreasing in the interval $(\lambda,+\infty)$, $F'(X)\psi(X)/(\mu\psi(X)-D)<-\infty$ as $X>\lambda$, a contradiction. Then there exists a neighborhood of λ, denoted by U, such that

$$\left(\frac{dV}{dt}\right)_{(3)}=(F(X)-F(\lambda))(\mu\psi(X)-D)>0, \tag{5}$$

as $X\in U\backslash\{\lambda\}$. $P(\lambda, g^{-1}(F(\lambda)))$ is an unstable singular point of (3). If (3) has limit cycles in the domain Ω, one of them is closest to the singular point P, which is inside stable and is denoted by Γ_1.

Let Γ be any closed orbit of (3); then

$$\oint_{\Gamma}\operatorname{div}(G, H)\,dt=\oint_{\Gamma}\psi'(X)(F(X)-g(Y))\,dt$$
$$+\oint_{\Gamma}\psi(X)F'(X)\,dt+\oint_{\Gamma}g'(Y)(\mu\psi(X)-D))\,dt.$$

Since

$$\oint_{\Gamma}\psi'(X)(F(X)-g(Y))\,dt=\oint_{\Gamma}(\psi'(X)/\psi(X))\,dX=0,$$
$$\oint_{\Gamma}g'(Y)(\mu\psi(X)-D)\,dt=\oint_{\Gamma}(g'(Y)/g(Y))\,dY=0,$$

then

$$\oint_{\Gamma}\operatorname{div}(G, H)\,dt=\oint_{\Gamma}\psi(X)F'(X)\,dt. \tag{6}$$

The X-coordinates of leftmost point M and rightmost point N at the orbit Γ_1 are denoted by $X_M=\min_{(X,Y)\in\Gamma_1}X$, $X_N=\max_{(X,Y)\in\Gamma_1}X$, respectively (see Fig. 1). Define

$$f_1(X)=F'(X)-\frac{F'(X_M)\psi(X_M)(\mu\psi(X)-D)}{(\mu\psi(X_M)-D)\psi(X)}.$$

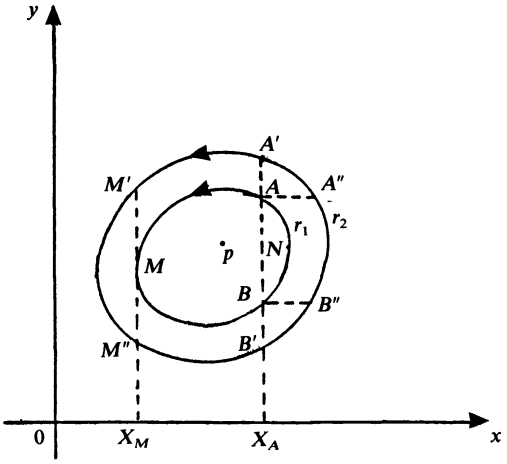


FIG. 1

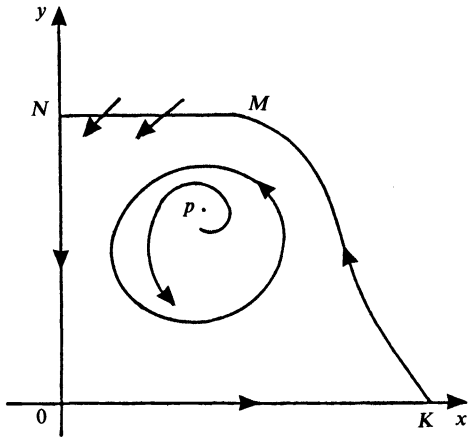


FIG. 2

$f_1(X_M)=0$, $f_1(X)\psi(X)/(\mu\psi(X)-D)$ is nonincreasing in the intervals $(0,\lambda)$ and $(\lambda,+\infty)$, the same as for $F'(X)\psi(X)/(\mu\psi(X)-D)$, therefore $f_1(X)<0$ as $X\in(0,X_M)$ and $f_1(X)>0$ as $X\in(X_M,\lambda)$. For any closed orbit Γ of (3) we have

$$\oint_\Gamma \operatorname{div}(G,H)\,dt=\oint_\Gamma \psi(X)f_1(X)\,dt \tag{7}$$

where $f_1(\lambda)=F'(\lambda)>0$. If $f_1(X)$ is not less than zero in the interval (λ,X_N) also, then

$$\oint_{\Gamma_1} \operatorname{div}(G,H)\,dt>0.$$

On the contrary, the closed orbit Γ_1 is inside stable. Hence there must exist an X_A, $\lambda<X_A<X_N$, such that $f_1(X)>0$ as $X\in(X_M,X_A)$, $f_1(X)<0$ as $X\in(0,X_M)\cup(X_A,+\infty)$.

There exists another limit cycle that is outside and closest to the limit cycle Γ_1, denoted by Γ_2. The vertical line through the point M intersects the orbit Γ_2 at the point M' and M'' (see Fig. 1). The horizontal line $X=X_A$ intersects the orbit Γ_2 at the point A' and B'. The horizontal line through the point A intersects the orbit Γ_2 at the point A''; the horizontal line through the point B intersects the orbit Γ_2 at the point B'' (see Fig. 1). We have

$$\oint_{\Gamma_1}\psi(X)f_1(X)\,dt=\left(\int_{\widehat{AM}}+\int_{\widehat{MB}}+\int_{\widehat{BNA}}\right)\psi(X)f_1(X)\,dt, \tag{8}$$

$$\oint_{\Gamma_2}\psi(X)f_1(X)\,dt=\left(\int_{\widehat{A'M'}}+\int_{\widehat{M'M''}}+\int_{\widehat{M''B'}}+\int_{\widehat{B'B''}}+\int_{\widehat{B''A''}}+\int_{\widehat{A''A'}}\right)\psi(X)f_1(X)\,dt. \tag{9}$$

Let $Y=Y_1(X)$ and $Y=Y_2(X)$ denote the functions of curves $\widehat{AM}$ and $\widehat{A'M'}$, respectively; then

$$\begin{aligned}&\int_{\widehat{A'M'}}\psi(X)f_1(X)\,dt-\int_{\widehat{AM}}\psi(X)f_1(X)\,dt\\ &\quad=-\left(\int_{X_M}^{X_A}\frac{f_1(X)\,dx}{F(X)-g(Y_1(X))}-\int_{X_M}^{X_A}\frac{f_1(X)\,dX}{F(X)-g(Y_2(X))}\right)\end{aligned} \tag{10}$$

$$= -\int_{X_M}^{X_A} \frac{f_1(X)(g(Y_2(X)) - g(Y_1(X)))}{(F(X) - g(Y_1(X)))(F(X) - g(Y_2(X)))} < 0.$$

Similarly, we can prove

$$\int_{\widehat{M''B'}} \psi(X) f_1(X)\, dt - \int_{\widehat{MB}} \psi(X) f_1(X)\, dt < 0. \tag{10'}$$

Since $f_1(X) > 0$ as $X \in (X_M, X_A)$, $f_1(X) < 0$ as $X \in (0, X_M) \cup (X_A, +\infty)$, we have

$$\int_{\widehat{M'M''}} \psi(X) f_1(X)\, dt = -\int_{Y_{M'}}^{Y_{M''}} \frac{\psi(X) f_1(X)\, dy}{h(Y)(\mu\psi(X) - D)} < 0, \tag{11}$$

$$\int_{\widehat{B'M''}} \psi(X) f_1(X)\, dt = \int_{X_{B'}}^{X_{B''}} \frac{f_1(X)\, dx}{F(X) - g(Y)} < 0, \tag{11'}$$

$$\int_{\widehat{A''A'}} \psi(X) f_1(X)\, dt < 0 \tag{11''}$$

where (X, Y) in every integral are the coordinates of the points at corresponding curves; $Y_{M'}$ and $Y_{M''}$ are Y-coordinates of the points M' and M'', respectively; and $X_{B'}$ and $X_{B''}$ are X-coordinates of the points B' and B''. Let $X = X_1(Y)$ and $X = X_2(Y)$ denote the functions of the curves $\widehat{BA}$ and $\widehat{B''A''}$, respectively; then

$$\begin{aligned} &\int_{\widehat{B''A''}} \psi(X) f_1(X)\, dt - \int_{\widehat{BA}} \psi(X) f_1(X)\, dt \\ &= \int_{Y_B}^{Y_A} \frac{\psi(X_2(Y)) f_1(X_2(Y))\, dY}{h(Y)(\mu\psi(X_2(Y)) - D)} - \int_{Y_B}^{Y_A} \frac{\psi(X_1(Y)) f_1(X_1(Y))\, dY}{h(Y)(\mu\psi(X_1(Y)) - D)} < 0. \end{aligned} \tag{12}$$

Combining (8)-(12), we have

$$\oint_{\Gamma_2} \operatorname{div}(G, H)\, dt - \oint_{\Gamma_1} \operatorname{div}(G, H)\, dt < 0. \tag{13}$$

Since

$$\oint_{\Gamma_1} \operatorname{div}(G, H)\, dt \leqq 0, \tag{14}$$

then

$$\oint_{\Gamma_2} \operatorname{div}(G, H)\, dt < 0. \tag{15}$$

If Γ_1 is a stable limit cycle, then Γ_1 and Γ_2 are both stable limit cycles, a contradiction. Now assume Γ_1 is a semistable limit cycle. Define

$$H(X) = \int_{\lambda}^{X} \frac{(\mu\psi(\xi) - D)\, d\xi}{\psi(\xi)} \quad \text{as } X > 0$$

and consider a new equation:

$$\dot{X} = \psi(X)(\bar{F}(X) - g(Y)), \qquad \dot{Y} = h(Y)(\mu\psi(X) - D) \tag{16}$$

where

$$\bar{F}(X) = \begin{cases} F(X) & \text{as } 0 < X < \lambda, \\ F(X) - \alpha H(X) & \text{as } \lambda < X, \quad 0 \leqq \alpha \ll 1. \end{cases}$$

As α varies, (16) is a rotated vector field and it is the equation (3) when $\alpha = 0$. According to the theory of the rotated vector field [1], as $0 < \alpha \ll 1$ the semistable limit cycle Γ_1

is broken down to two limit cycles Γ_1' and Γ_1'', Γ_1' is on the inside of Γ_1'', Γ_1' is a stable limit cycle, and Γ_1'' is an unstable limit cycle. From (13) we have

$$\oint_{\Gamma_1''} \operatorname{div}(G, H)\, dt - \oint_{\Gamma_1'} \operatorname{div}(G, H)\, dt < 0,$$

a contradiction. After the discussion above we conclude that (3) has at most one stable limit cycle in the domain Ω. The proof is complete.

Remark 1. If $(\lambda, g^{-1}(F(\lambda)))$ is an unstable singular point in the domain Ω, i.e., equation $\mu\psi(X) - D = 0$ has a positive solution λ, K is the smallest positive solution of the equation $\varphi(X) = 0$, $K > \lambda$ and $\varphi(X) > 0$ as $X \in (0, K)$, then (3) has at least one limit cycle in the domain Ω. The point $K(K, 0)$ is a saddle-type singular point in the domain $\bar{\Omega} = \{(X, Y) \mid X \geqq 0, Y \geqq 0\}$. If the orbit of (3) from the saddle point in the domain $\bar{\Omega}$ does not approach infinity, there exists the highest point M at this orbit (see Fig. 2). The horizontal line from the point M intersects the X-axis at the point N. Then if the closed curve $\widehat{KMNOK}$ surrounds a Poincaré-Bendixon region, by the Poincaré-Bendixon theorem, equation (3) has at least one stable limit cycle in the domain Ω. Adding the conditions $K > \lambda$ and $\varphi(X) > 0$ as $X \in (0, K)$ in Theorem 1, we have that (3) has exactly one stable limit cycle in the domain Ω.

Remark 2. If $F'(X)\psi(X)/(\mu\psi(X) - D)$ is nonincreasing in the intervals $(0, \lambda)$ and (λ, K), K is mentioned in Remark 1, and $K > \lambda$, the conclusion of Theorem 1 also holds.

THEOREM 2. *If all the positive solutions X, Y of the system of equations*

$$H(X) = H(Y), \qquad F(X) = F(Y) \tag{17}$$

satisfy $X = Y$, then equation (3) *does not have a periodic solution in the domain Ω.*

Proof. Since $\psi(0) = 0$ and $\psi'(X) > 0$, then $\psi(X) > 0$ as $X > 0$. Assume λ is the unique positive solution of the equation $\mu\psi(X) - D = 0$ in the interval $(0, +\infty)$. $H(X)$ is monotone decreasing in the interval $(0, \lambda)$ and monotone increasing in the interval $(\lambda, +\infty)$. $H(X)$ has inverse functions in the intervals $(0, \lambda)$ and $(\lambda, +\infty)$, denoted by $X_2(H)$ and $X_1(H)$, respectively. Suppose (3) has a periodic solution, i.e., (3) has a closed orbit surrounding the singular point P, denoted by Γ. Let the equations of the curve Γ be

$$X = \Gamma_1(Y) \quad \text{as } \lambda \leqq X,$$

$$Y = \Gamma_2(Y) \quad \text{as } 0 < X \leqq \lambda.$$

There exist $Y_{\min}$ and $Y_{\max}$, a Y-coordinate of the lowest point and a Y-coordinate of the highest point at the orbit Γ, respectively, and from (3) there hold

$$\frac{dH(\Gamma_1(Y))}{dY} = \frac{F(X_1(H(\Gamma_1(Y)))) - g(Y)}{h(Y)}, \tag{18}$$

$$\frac{dH(\Gamma_2(Y))}{dY} = \frac{F(X_2(H(\Gamma_2(Y)))) - g(Y)}{h(Y)} \tag{18'}$$

where $Y_{\min} \leqq Y \leqq Y_{\max}$. From (18) and (18'), the curves $\Gamma_1(Y)$ and $\Gamma_2(Y)$ are the solutions of the boundary value problems

$$\frac{d\bar{H}}{dY} = \frac{F(X_1(\bar{H})) - g(Y)}{h(Y)}, \qquad \bar{H}(Y_{\min}) = \bar{H}(Y_{\max}) = 0 \tag{19}$$

and

$$\frac{d\bar{H}}{dY}=\frac{F(X_2(\bar{H}))-g(Y)}{h(Y)}, \qquad \bar{H}(Y_{\min})=\bar{H}(Y_{\max})=0, \tag{19'}$$

respectively. We will prove that $F(X_1(\bar{H}))\not\equiv F(X_2(\bar{H}))$ as $\bar{H}>0$. Without loss of generality, assume $F(X_1(\bar{H}))<F(X_2(\bar{H}))$ as $\bar{H}>0$; then

$$\frac{F(X_1(\bar{H}))-g(Y)}{h(Y)}<\frac{F(X_2(\bar{H}))-g(Y)}{h(Y)} \quad \text{as } \bar{H}>0. \tag{20}$$

If (19) and (19') both have solutions denoted by $\bar{H}_1(Y)$ and $\bar{H}_2(Y)$, respectively, then there exists $0<\varepsilon\ll 1$, such that $\bar{H}_1(Y)<\bar{H}_2(Y)$ when $Y_{\min}<Y<Y_{\min}+\varepsilon$ and $\bar{H}_1(Y)>\bar{H}_2(Y)$ when $Y_{\max}-\varepsilon<Y<Y_{\max}$. Then there exists Y_0, $Y_{\min}+\varepsilon<Y_0<Y_{\max}-\varepsilon$, such that $\bar{H}_1(Y_0)=\bar{H}_2(Y_0)$ and $d\bar{H}_1(Y_0)/dY\geqq d\bar{H}_2(Y_0)/dY$, a contradiction. Now we prove $F(X_1(\bar{H}))\not\equiv F(X_2(\bar{H}))$ as $\bar{H}>0$. If there is an $\bar{H}_0>0$ such that $F(X_1(\bar{H}_0))=F(X_2(\bar{H}_0))$, and if we let $X_0=H_1(\bar{H}_0)$ and $Y_0=X_2(\bar{H}_0)$, then $F(X_0)=F(Y_0)$, $H(X_0)=H(Y_0)$ and $X_0>Y_0>0$, which contradicts the condition of the theorem. So $F(X_1(\bar{H}))\not\equiv F(X_2(\bar{H}))$ as $\bar{H}>0$. The theorem is proved.

Now we use Theorems 1 and 2 to investigate the predator-prey system (1).

THEOREM 3. *There are three types of global structure of the predator-prey system* (1), *in which* (1)

(a) *Has an unique stable limit cycle in the domain* Ω, *and all the trajectories of* (1) *in the domain* Ω *approach this limit cycle*;

(b) *Has a stable singular point* $P(\lambda, g^{-1}(F(\lambda)))$ *in the domain* Ω, *and all the trajectories of* (1) *in the domain* Ω *terminate at the singular point*;

(c) *Does not have any singular point in the domain* Ω, *and all trajectories of* (1) *in the domain* Ω *terminate at the point* $K(K,0)$.

The types of global structure of (1) *under the various conditions of the parameters* γ, a, μ, K, *and* D *are*:

when $n=1$

Distribution of parameters	$\mu>D>0$		$D\geqq\mu>0$
	$K>(\mu+D)a/(\mu-D)$	$K\leqq(\mu+D)a/(\mu-D)$	
Types of structure	type (a)	type (b)	type (c)

when $n=2$

Distribution of parameters		$2D>\mu>D>0$		
	$\mu\geqq 2D>0$	$K>K^*$	$K\leqq K^*$	$D\geqq\mu>0$
Types of structure	type (b)	type (a)	type (b)	type (c)

where $K^*=2D\sqrt{aD/(\mu-D)}/(2D-\mu)$.

Before we prove Theorem 3 we give the following lemma.

LEMMA 4. *If* $0<Y\leqq X$, *then*

$$\frac{X-Y}{X+Y}\leqq\frac{1}{2}(\ln X-\ln Y). \tag{21}$$

The equality holds if and only if $X=Y$.

Proof. Assume $S \geqq 1$; then $(S-1)^2 \geqq 0$, and moreover $(S+1)^2 \geqq 4S$. Therefore $2/(S+1)^2 \leqq 1/2S$, and integrating, we have

$$\int_1^r \frac{2dS}{(S+1)^2} \leqq \int_1^r \frac{1}{2}\frac{dS}{S}, \qquad 1-\frac{2}{r+1} \leqq \frac{1}{2}\ln r. \tag{22}$$

Let $r = X/Y$; from (22) we obtain (21). If $X = Y$, then the equality in (21) holds. Inversely, if the equality in (22) holds, then $r = 1$, i.e., $X = Y$. The lemma is proved.

Proof of Theorem 3.

(i) When $n = 1$, $F(X) = \gamma(1 - X/K)(a + X)$. As $K > a + 2\lambda = a(\mu + D)/(\mu - D)$, $0 < D < \mu$, then

$$KF'(\lambda) = \gamma(-2\lambda + K - a) > 0.$$

Let

$$\begin{aligned}\Delta(X) &= F'(X)\psi(X)/(\mu\psi(X) - D) \\ &= \gamma(-2X^2 + (K-a)X)/K((\mu - D)X - Da);\end{aligned}$$

then

$$\Delta'(X) = \frac{-2(\mu - D)X^2 + 4DaX - (K-a)Da}{K((\mu - D) - Da)^2}.$$

The determinant of $-2(\mu - D)X^2 + 4DaX - (K-a)Da$ is $8Da(\mu - d) \times (2\lambda + a - K) < 0$. Hence $\Delta'(X) < 0$ as $X \in (0, +\infty)\backslash\{\lambda\}$, i.e., $\Delta(X)$ is nonincreasing in the intervals $(0, \lambda)$ and $(\lambda, +\infty)$. Using Theorem 1 and Remark 1, when $n = 1$ and $K > a + 2\lambda$, we have that (1) has exactly one stable limit cycle in the domain Ω. The global structure in this case is of type (a).

(ii) When $n = 1$, $\mu > D > 0$, $K \leqq a + 2\lambda = a(\mu + D)/(\mu - D)$, then

$$H(X) = (\mu - D)(X - \lambda) - Da(\ln X - \ln \lambda).$$

Now (17) is

$$\begin{aligned}&(\mu - D)(X - Y) - Da(\ln X - \ln Y) = 0, \\ &(K - a)(X - Y) - (X^2 - Y^2) = 0.\end{aligned} \tag{23}$$

If the system of equations has positive solutions X^*, Y^*, satisfying $X^* \neq Y^*$, without loss of generality, assume $0 < X^* < Y^*$, then we have

$$\begin{aligned}&(\mu - D)(X^* - Y^*) - Da(\ln X^* - \ln Y^*) = 0, \\ &(K - a)(X^* - Y^*) - (X^* + Y^*)(X^* - Y^*) = 0.\end{aligned} \tag{24}$$

Since $K \leqq (\mu + D)a/(\mu - D)$, from (23) there is

$$\begin{aligned}\frac{Y^* - X^*}{Y^* + X^*} &= \frac{Da}{(\mu - D)(K - a)}(\ln Y^* - \ln X^*) \\ &\geqq \frac{1}{2}(\ln Y^* - \ln X^*),\end{aligned}$$

a contradiction. Hence the system of equations (23) has no such positive solutions X^*, Y^*, satisfying $X^* \neq Y^*$. Therefore by Theorem 2, (1) has no closed trajectory in the domain Ω. The singular point $P(\lambda, g^{-1}(F(\lambda)))$ in this case is stable, and all the trajectories of (1) terminate at this singular point. The global structure of (1) in this case is of type (b).

(iii) When $n=1$, $\mu \leqq D$, there is no singular point in the domain Ω for (1). In the domain Ω, $K(K,0)$ is a stable node-type singular point and $O(0,0)$ is a saddle-type singular point, and all the trajectories of (1) in the domain Ω terminate at the point $K(K,0)$. The global structure in this case is of type (c).

(iv) When $n=2$, $0<D<\mu<2D$, $K>K^*=2D\lambda/(2D-\mu)$,

$$F(X)=\gamma(1-X/K)(a+X^2)/X,$$

$$\lambda^2F'(\lambda)=\gamma(-a+\lambda^2-3\lambda^3/K)$$

$$=\gamma a(\mu-D)(2D-\mu-2D\lambda/K)>0,$$

$$\Delta(X)=F'(X)\psi(X)/(\mu\psi(X)-D)$$

$$=\gamma(-a+X^2-3X^3/K)/((\mu-D)X^2-Da),$$

$$\Delta'(X)=\frac{2\gamma(((\mu-D)a-Da)+3DaX/K-(\mu-D)X^3/K)X}{((\mu-D)X^2-Da)^2}.$$

Since

$$\lambda^2K(((\mu-D)a-Da)+3DaX/K-(\mu-D)X^3/K)<-Da(X-\lambda)^2(X+2\lambda)<0,$$

as $X\in(0,\lambda)\cup(\lambda,+\infty)$, then $\Delta'(X)<0$, as $X\in(0,+\infty)\backslash\{\lambda\}$, i.e., $\Delta(X)$ is nonincreasing in the intervals $(0,\lambda)$ and $(\lambda,+\infty)$. According to Theorem 1 and Remark 1, (1) has exactly one stable limit cycle in the domain Ω as $n=2$ and $K>K^*$. The global structure of (1) in this case is of type (a).

(v) When $n=2$, $0<D<\mu<2D$, $K\leqq K^*$, the system (17) is

$$(1-X/K)(a+X^2)/X=(1-Y/K)(a+Y^2)/Y,$$

$$(\mu-D)(X-\lambda)+Da\left(\frac{1}{X}-\frac{1}{\lambda}\right)=(\mu-D)(Y-\lambda)+Da\left(\frac{1}{Y}-\frac{1}{\lambda}\right). \tag{25}$$

The system of equations has no such positive solutions X^*, Y^*, $X^*\neq Y^*$. On the contrary, if (25) has such solutions, we have

$$-a+X^*Y^*-(X^*+Y^*)X^*Y^*/K=0,$$

$$X^*Y^*=\lambda^2, \tag{26}$$

$$X^*+Y^*=(\lambda^2-a)K/\lambda^2=(2D-\mu)K/D$$

$$\leqq 2\sqrt{\frac{2D}{\mu-D}}=2\lambda, \tag{27}$$

$$2\lambda=2\sqrt{X^*Y^*}<(X^*+Y^*)\leqq 2\lambda,$$

a contradiction. By Theorem 2, (1) has no closed trajectory in the domain Ω, and all the trajectories of (1) in the domain Ω terminate at the stable singular point $P(\lambda, g^{-1}(F(\lambda)))$. The global structure of (1) in this case is of type (b).

(vi) When $n=2$, $\mu \geqq 2D>0$, similar to case (v), we have systems of equations (25) and (26). From (26), we have

$$0<X^*+Y^*=(2D-\mu)K/D \leqq 0,$$

a contradiction, and the same result as in case (v) holds.

(vii) When $n=2$, $0 \geqq \mu > D$, similar to case (iii), the same result holds.

Combining (i)-(vii) completes the proof.

We give Figs. 3-5 to illustrate the types of the structure (a), (b), and (c), respectively.

In Fig. 3, type (a), the system (1) finally approaches a closed orbit. The numbers of predator and prey both vary periodically.

In Fig. 4, type (b), the system (1) finally terminates at a unique stable equilibrium, i.e., terminates at the stable singular point $P(\lambda, g^{-1}(F(\lambda)))$. The numbers of predator and prey limit to nonzero constants.

In Fig. 5, type (c), the birth rate of the predator is less than its death rate, i.e.,

$$\dot{Y}=Y(\mu X^n/(a+X^n)-D)<0.$$

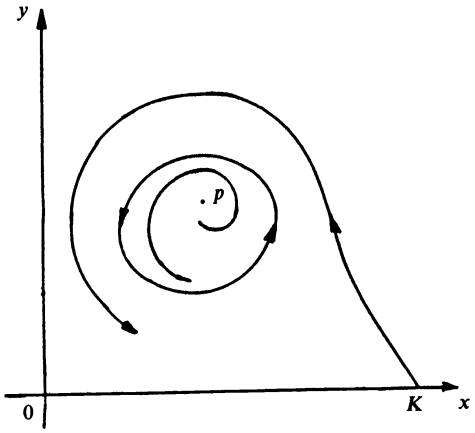


FIG. 3

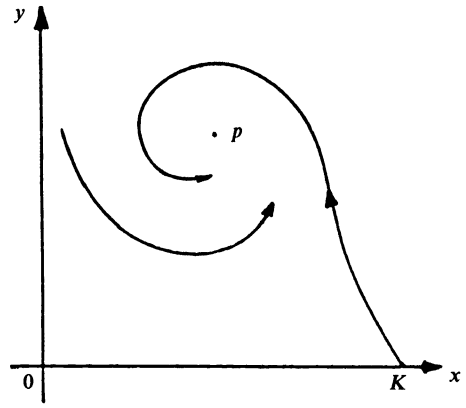


FIG. 4

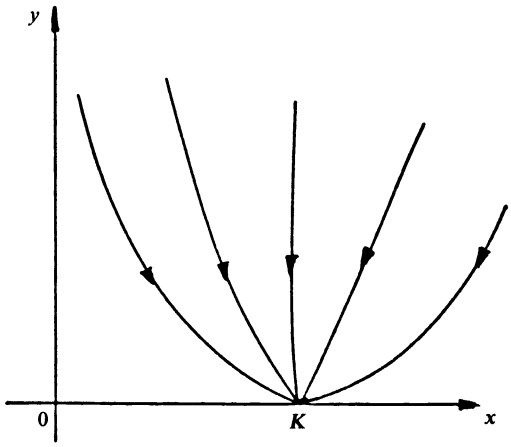


FIG. 5

The number of predators is monotonic decreasing, and finally terminates at zero. The number of prey will be constant, $\dot{X} = X(1 - X/K) = 0$ as $X = K$. The predator-prey system (1) will be exterminated.

Acknowledgment. We thank a referee who pointed out a recent paper of Kuang and Kreedman who obtained similar results using a different method.

REFERENCES

[1] G. F. D. DUFF, *Limit cycle and rotated vector fields*, Ann. of Math. (2), 57 (1953), pp. 15-31.
[2] A. HASTINGS, *Global stability of two species systems*, J. Math. Biol., 5 (1978), pp. 599-603.
[3] S. B. HSU, S. P. HUBBELL, AND P. WALTMAN, *Competing predators*, SIAM J. Appl. Math., 35 (1978), pp. 617-625.
[4] D. W. JORDAN, AND P. SMITH, *Nonlinear Ordinary Differential Equations*, Clarendon Press, Oxford, 1977.
[5] C. KUO-SHENG, *Uniqueness of limit cycle for a predator-prey system*, SIAM J. Math. Anal., 12 (1981), pp. 541-548.
[6] S. LEFSCHETZ, *Ordinary Differential Equations: Geometric Theory*, Interscience, New York, 1975.
[7] D. SUNHONG, *Uniqueness of limit cycle of predator-prey system*, Kezue Tongbao, 13 (1985), pp. 785-788. (In Chinese.)
[8] K. YANG AND H. I. KREEDMAN, *Uniqueness of limit cycle in Gauss-type models of predator-prey systems*, Math. Biosci., 88 (1988), pp. 67-84.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1436-1446, November 1989

009

THE UNIQUENESS AND STABILITY OF THE REST STATE FOR STRONGLY COUPLED OSCILLATORS*

G. BARD ERMENTROUT† AND WILLIAM C. TROY†

Abstract. In this paper it is shown that an unstable steady state can be stabilized in the presence of sufficient inhomogeneity and strong diffusion for a continuum and discrete model. These results are applied to a pair of coupled oscillators and to an oscillatory reaction-diffusion system. It is also shown that for the reaction-diffusion system, the trivial rest state is the unique phase-locked solution. Some additional numerical results are presented that illustrate the nature of this bifurcation for a realistic model oscillator.

Key words. oscillation, stability, reaction-diffusion equations

AMS(MOS) subject classifications. 35B35, 58F10

1. Introduction. This paper is a continuation of an analysis of an oscillatory reaction-diffusion system in the presence of a gradient in natural frequencies. In our previous article [1], we have shown the existence of a periodic (phase-locked) solution and shown some of its behavior via numerical methods. In this paper, we are interested in the stabilization and uniqueness of the rest state (which is unstable in the absence of the diffusion). This phenomenon of restabilization has been explored numerically for the Brusselator by Bar-Eli [2] and has been shown for a simple λ-ω model in [3]. In both of the latter cases, two discrete oscillators are coupled diffusively and it is shown that if the two uncoupled frequencies are different enough, then there is a stable equilibrium for coupling strengths in some intermediate range. More recently, a different mechanism for the stabilization of a rest state (and thus, the death of the oscillation) was discussed in [4] for *identical* oscillators coupled *directly* rather than diffusively. In a sense, this is the inverse of Smale's mechanism by which two coupled "dead" cells could spontaneously begin to oscillate if allowed to interact through diffusion. In the Smale article [5], though, the two cells are identical and the mechanism is more akin to a Turing instability of the Hopf type (see, e.g., [6]).

In this paper, we are mainly interested in a continuum of cells with a linear gradient of frequencies that are coupled by ordinary diffusion. We prove the stability and uniqueness of the stationary trivial solution for sufficiently strong diffusion and a large enough gradient in frequencies. We will also consider a general pair of coupled oscillators in $\mathbb{R}^n$ and show how the trivial state can be restabilized if the two are different enough and the coupling is sufficiently strong. In § 2, we describe the discrete model and prove the stability result. Section 3 introduces the continuum model and the stability theorem is stated and proved. Section 4 is devoted to a proof of the uniqueness of the trivial solution as a phase-locked state.

2. Two coupled oscillators. Consider the following pair of coupled nonlinear differential equations:

$$\frac{du}{dt} = F(u) + \beta(v-u), \tag{2.1}$$

$$\frac{dv}{dt} = \sigma F(v) + \beta(u-v) \tag{2.2}$$

* Received by the editors April 24, 1987; accepted for publication (in revised form) January 30, 1989.
† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.
This work was supported by National Science Foundation contract DMS 87-01405.

where β is a scalar representing the strength of coupling and σ is a parameter that distinguishes the rates of the two systems. In particular, when $\beta = 0$ we may suppose that each system oscillates about some rest point which we may assume to be zero with no loss in generality. Then the ratio of the frequencies of the two oscillators is σ. We assume that in the absence of coupling, the equilibrium point $u = v = 0$ is unstable. It will remain so for sufficiently small strengths of coupling. We now show how the nature of this instability determines whether it is always unstable for all choices of (β, σ). In fact, we will show that if the eigenvalues of the linearized uncoupled system are complex and the imaginary part is large enough, there are ranges of parameters β and σ such that the coupled system is linearly stable.

Let A be the Jacobian of F evaluated at $u = 0$. Consider the following linear system:

$$\frac{du}{dt} = Au + \beta(v - u), \tag{2.3}$$

$$\frac{dv}{dt} = \sigma Av + \beta(u - v). \tag{2.4}$$

LEMMA 2.1. *If (ν, ϕ) is an eigenvalue-eigenvector pair for the matrix A, then (γ, ξ) is an eigenvalue-eigenvector pair for the matrix*

$$M = \begin{bmatrix} A - \beta I & \beta I \\ \beta I & \sigma A - \beta I \end{bmatrix}$$

where $\xi = (\phi, s\phi)^T$, γ satisfies

$$\gamma^2 + (2\beta - \nu(\sigma + 1))\gamma + \nu(\sigma\nu - \beta(\sigma + 1)) = 0 \tag{2.5}$$

and $(1, s)^T$ is the eigenvector corresponding to γ for the 2×2 matrix

$$\begin{bmatrix} \nu - \beta & \beta \\ \beta & \sigma\nu - \beta \end{bmatrix}.$$

Proof. Substitute the vector ξ into the eigenvalue equation for M and use the fact that $A\phi = \nu\phi$. □

LEMMA 2.2. *Let $\nu = \lambda + i\omega$ denote an eigenvalue of A.*

(i) *If $\omega = 0$ and $\lambda > 0$, then $(0, 0)$ is unstable as a solution to* (2.3).

(ii) *If $\lambda < 0$, then the corresponding pair of eigenvalues for the linear system* (2.3)-(2.4) *have negative real parts.*

Proof. We need only analyze the roots of (2.5).

(i) If ν is real and positive, then for stability, we must have the coefficients of (2.5) both positive, i.e.,

$$\frac{\sigma}{\sigma + 1} > \frac{\beta}{\nu} > \sigma + 1.$$

Since the first expression is less than 1 and the last is greater, this is an obvious impossibility.

(ii) If ν is real and negative, then the coefficients of (2.5) are both positive so that all the roots have negative real parts. Suppose that ν is complex with a negative real part. We will prove that it is impossible for eigenvalues of (2.5) to cross the imaginary axis. When $\beta = 0$, the eigenvalues are ν and $\sigma\nu$, both having negative real parts. We follow these eigenvalues as β increases. There are two ways in which instability could occur; one of the two eigenvalues of (2.5) could become real and cross through zero, or it could cross into the right halfplane through the imaginary

axis. Since ν is complex, the former is impossible; there will be a zero eigenvalue if and only if $\beta = \sigma\nu/(\sigma+1)$, which is complex. Thus, if instabilities are to occur, they must be through the imaginary axis. We differentiate (2.5) with respect to β and find that

$$\frac{d\gamma}{d\beta} = \frac{-|\nu(\sigma-1)-2\gamma|^2+2\beta[\nu(\sigma+1)-2\gamma]}{|-\nu(\sigma+1)+2\gamma+2\beta|^2}. \tag{2.6}$$

Suppose that for some β, $\gamma = i\kappa$ (i.e., $\mathrm{Re}\,(\gamma)=0$). Then, the real part of $d\gamma/d\beta$ is strictly negative since $\mathrm{Re}\,(\nu)=\lambda<0$. Thus, it is impossible to cross the imaginary axis, and the roots to (2.5) lie in the left-half complex plane. □

Lemma 2.2 takes care of the case of real eigenvalues and complex eigenvalues with negative real parts. The only remaining case is the most interesting from our point of view: the case of complex eigenvalues with positive real parts. The proof of (ii) depended crucially on the fact that $\mathrm{Re}\,(\nu)<0$; the proof cannot be pushed through to show that eigenvalues remain on the right halfplane for all β. In fact, it is clear from (2.6) that for small β, $d\gamma/d\beta<0$ so that the eigenvalues move toward the left. But, as β gets larger, (2.6) may become positive again. As $\beta\to\infty$, one of the roots to (2.5) tends to $\frac{1}{2}\nu(\sigma+1)$, which has a positive real part. Thus, it is not obvious that eigenvalues can be pushed to the left halfplane.

LEMMA 2.3. *Let $q=\lambda/\omega$, where $\nu=\lambda+i\omega$. If q is sufficiently small, then there are regimes of β and σ such that the roots of* (2.5) *have negative real parts.*

Proof. The solutions to (2.5) are

$$\gamma = -\beta+\nu(\varepsilon+1)\pm\sqrt{\beta^2+\varepsilon^2\nu^2} \tag{2.7}$$

where $\varepsilon=(\sigma-1)/2$. Choose $\beta\approx\beta^*=|\varepsilon\omega|\sqrt{1-q^2}$ and substitute into (2.7). Using the fact that $\nu^2=\omega^2(q^2+2iq-1)$, we see that the maximal real part (corresponding to the +) is approximately

$$\mathrm{Re}\,\gamma \approx \omega[-\varepsilon\sqrt{1-q^2}+q(\varepsilon+1)+\varepsilon\sqrt{q}]. \tag{2.8}$$

For q sufficiently small and $\varepsilon>0$ (corresponding to $\sigma>1$, which may be assumed with no loss of generality), it is clear that $\mathrm{Re}\,\gamma<0$ as required. For β near β^*, the real part of the eigenvalue stays negative. □

Remark. It is clearly necessary that $\sigma\neq 1$ to obtain restabilization for any value of β. Indeed, setting $\varepsilon=0$ ($\sigma=1$) in (2.8) yields an eigenvalue ν that has a positive real part. The asymmetry in the problem is absolutely necessary for the coupling to change the signature of the eigenvalues. Furthermore, the greater the asymmetry, the larger q may be. The largest allowable q is roughly 0.35; for q greater than this, (2.8) is always positive for all $\varepsilon\geqq 0$. We also note that this is an intermediate value of coupling strength, β, neither very large, nor very small.

The following proposition follows from the previous lemmas.

PROPOSITION 2.1. *Suppose that zero is an unstable equilibrium point for* (2.3) *or* (2.4) *when $\beta=0$. Assume the instability is due to a single pair of complex conjugate eigenvalues $\lambda\pm i\omega$. Then, if λ/ω is small enough, these are values of σ and β for which the origin is an asymptotically stable equilibrium point.*

Remarks. The hypothesis on the signs and relative sizes of the eigenvalues is satisfied, for example, after a Hopf bifurcation.

In [3], we study a pair of symmetrically coupled oscillators and show that if the frequency difference between them is sufficiently large, then as the coupling increases, the amplitude of the phase-locked solution decreases and eventually goes to zero in a manner analogous to the continuum model considered in [1].

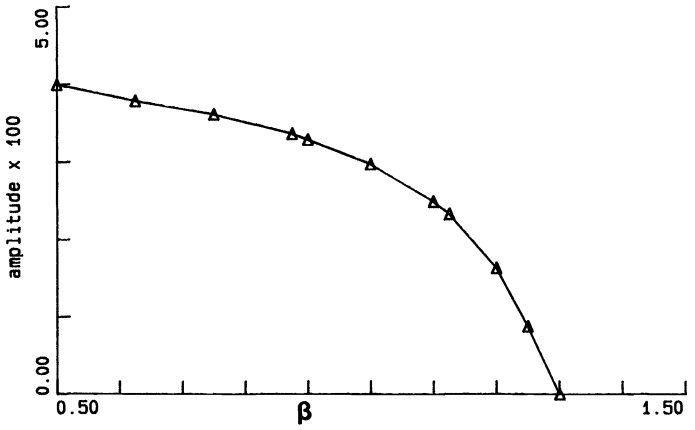


FIG. 1. *The difference between the maximum and minimum value of $u(t)$ as a function of the strength of diffusion for the Wilson-Cowan equations with $\sigma = 2$.*

Before continuing with the stability of the equilibrium for a continuous model, we describe some numerical results for a coupled oscillator model of the form (2.1)-(2.2). We consider a two-dimensional system, specifically the Wilson-Cowan equations [7]:

$$\frac{du}{dt} = -u + f(16.0u - 20.0v - 0.5), \tag{2.9}$$

$$\frac{dv}{dt} = -v + f(18.0u - 3.0) \tag{2.10}$$

where $f(u) = .5(1+\tanh(u))$. With these parameter values, the Wilson-Cowan equations oscillate about an unstable equilibrium point, $(u, v) = (0.110065, 0.115304)$. The eigenvalues of the Jacobian are $0.58045 \pm i\ 3.77018$, so that $q = 0.15396$. We have chosen $\sigma = 2$, so that the value $\beta^* = 1.862$. In Fig. 1, we have sketched the amplitude of $u(t)$ (i.e., the difference between the maximum and minimum values of $u(t)$) as a function of the parameter β for $\beta \in [0.5, 1.5]$. In this regime, the two oscillators are phase-locked in a 1:1 ratio. The amplitude gradually decreases until at a critical value of $\beta \approx 1.3$, the steady state is restabilized. This phenomenon is identical to that proved in [3] for a special solvable oscillator model. We have applied these results to a variety of other oscillator models and find similar behavior; if the difference in natural frequencies is too great, then for certain strengths of coupling, the rest state is restabilized as the amplitude of coupling decreases.

3. The continuum model. In [1], we have analyzed the phase-locked behavior of a simple reaction-diffusion equation with a linear frequency gradient:

$$\frac{\partial u}{\partial t} = \lambda(R)u - \omega(R, x)v + d\frac{\partial^2 u}{\partial x^2}, \tag{3.1}$$

$$\frac{\partial v}{\partial t} = \lambda(R)v + \omega(R, x)u + d\frac{\partial^2 v}{\partial x^2}, \tag{3.2}$$

$$\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = \frac{\partial v}{\partial x}(0, t) = \frac{\partial v}{\partial x}(1, t) = 0 \tag{3.3}$$

where $R=u^2+v^2$, $\omega(R,x)=\omega_0-\hat{\sigma}x+q(1-R)$, and $\lambda(R)=\Lambda(1-R)$. In absence of diffusion, (3.1)-(3.2) admit stable periodic solutions with frequency $\omega_0+\hat{\sigma}x$. In [1], we have shown that for d large enough and $q=0$, there are periodic (phase-locked) solutions to (3.1)-(3.3). We have also shown numerically that for fixed $d\geqq 1$, as $\hat{\sigma}$ increases, the maximum amplitude $\sqrt{R}$ decreases to zero. We also conjectured that for $\hat{\sigma}$ large enough, the rest state stabilized. In the next section, we prove that for sufficiently large $\hat{\sigma}$ the solution $u=v=0$ is the unique periodic solution to (3.1)-(3.3). In this section, we prove that $u(x,t)=v(x,t)=0$ is a linearly stable solution to (3.1)-(3.3).

It is convenient to introduce the complex variable $w(x,t)=u(x,t)+iv(x,t)$. The linear stability problem is then

$$\frac{\partial w}{\partial t}=[\Lambda+i(\omega_0-\hat{\sigma}x)]w+d\frac{\partial^2 w}{\partial x^2}, \tag{3.4}$$

$$\frac{\partial w}{\partial x}(0,t)=\frac{\partial w}{\partial x}(1,t)=0. \tag{3.5}$$

We seek solutions to (3.4)-(3.5) of the form $w(x,t)=\exp(-[\hat{\lambda}+i\omega_0]t)z(x)$. $z(x)$ satisfies the ordinary differential equation

$$dz''+(\Lambda-i\hat{\sigma}x)z=-\hat{\lambda}z, \tag{3.6}$$

$$z'(0)=z'(1)=0. \tag{3.7}$$

Divide through by d and let $c=\Lambda/d$, $\sigma=\hat{\sigma}/d$, $\lambda=\hat{\lambda}/d$. Equation (3.6) becomes

$$z''+(c-i\sigma x)z=-\lambda z. \tag{3.8}$$

Finally, we let $\lambda=\alpha+i\delta$, $z=re^{i\theta}$, $\psi=\theta'$, $s=r'/r$, so that (3.7), (3.8) become

$$s'=\psi^2-s^2-(\alpha+c), \tag{3.9}$$

$$\psi'=\sigma x-\delta-2s\psi, \tag{3.10}$$

$$s(0)=\psi(0)=0, \tag{3.11}$$

$$s(1)=\psi(1)=0. \tag{3.12}$$

The origin is stable if and only if solutions to (3.6)-(3.7) have solutions with $\mathrm{Re}\,(\hat{\lambda})>0$, for then $w(x,t)$ decays exponentially in time. Since $\alpha=\mathrm{Re}\,(\hat{\lambda})/d$, to prove the equilibrium is stable we must show that (3.9)-(3.12) has a solution if and only if $\alpha>0$. It is clear from (3.8) that when $\sigma=0$, (3.7)-(3.8) has a solution, $z\equiv C$, a constant, and $\lambda=-c<0$, so that the origin is unstable. Thus, it is crucial that $\sigma\neq 0$. We can assume with no loss of generality that $\sigma>0$; if $\sigma<0$, then let $x\to 1-x$ in (3.4)-(3.5) and let $\omega_0\to\omega_0+\sigma$. We will prove the following theorem.

THEOREM 3.1. *Let $0\leqq c\leqq 1$. There is a $\sigma^*>0$ such that if* (3.9)-(3.12) *has a solution for some $\sigma>\sigma^*$ then $\alpha>0$.*

Since the proof is somewhat technical in nature, we give a short outline of the proof. Our goal is to show that if σ is large and (3.9)-(3.12) has a solution, then α must be positive. Thus, for the sake of contradiction, we assume that for every large σ there is an $\alpha\equiv\alpha(\sigma)\leqq 0$ for which (3.9)-(3.12) does indeed have a solution. The first step in our analysis is to prove that any solution to (3.9)-(3.12) must satisfy $|s(x)|\leqq\tan(1)$ for all $x\in(0,1)$. Next, we determine that if (3.9)-(3.12) holds then δ, α, and σ must lie in the parameter range $\alpha\geqq -c$ and $\sigma\geqq\delta>0$. We then define a σ^* independent of α, c, and δ, and assume that a solution exists for some $\sigma>\sigma^*$, $\alpha\in[-c,0]$ and $\delta\in[0,\sigma]$. We break the remaining analysis into two subcases, $0\leqq\delta\leqq\sigma/2$ and

$\sigma/2 \leqq \delta \leqq \sigma$. For the first case, we consider the x-intervals $I_1 = [\frac{5}{8}, \frac{3}{4}]$ and $I_2 = [\frac{3}{4}, \frac{7}{8}]$. We need to show that $|s(x_0)| = \tan(1)$ at some $x_0 \in I_1 \cup I_2$ to arrive at a contradiction. The key to this is to analyze (3.10) and obtain a lower bound estimate on the function $|\psi|$. We first show that there is an $M > 0$ (independent of σ) such that if $|\psi| \geqq M$ at some point in I_1, then $|\psi| \geqq M$ for all $x \in I_2$. This lower bound on $|\psi|$ is then used in (3.9) to show that $|s|$ must exceed $\tan(1)$. The remainder of this subcase is therefore devoted to proving that for all large values of σ, $|\psi| \geqq M$ at some value in I_1. The second case, $\sigma/2 \leqq \delta \leqq \sigma$ is handled in the same fashion. Here, we analyze the behavior of ψ on $J_1 = [\frac{1}{8}, \frac{1}{4}]$ and $J_2 = [\frac{1}{4}, \frac{3}{8}]$.

LEMMA 3.1. *If* (1)-(4) *has a solution then* $|s| \leqq \tan(1)$ *for all* $x \in [0, 1]$.

Proof. From (3.9) we have $s' \geqq -(s^2+1)$ on $[0, 1]$. Thus, $ds/(s^2+1) \geqq -dx$ and so $\tan^{-1}(s(x)) \geqq \tan^{-1}(s(x_0)) - (x - x_0)$ for $x_0 \in [0, 1)$ and $x_0 \leqq x \leqq 1$. From this, we conclude that $\tan^{-1}(s(1)) \geqq \tan^{-1}(s(x_0)) - 1$. If $s(x_0) > \tan(1)$ for some $x_0 \in [0, 1)$, then $s(1) > 0$. This contradicts (3.12). □

Proof of Theorem 3.1. Suppose that a solution of (3.9)-(3.12) exists for some $\sigma > 0$, $\delta \in \mathbb{R}$, and $\alpha \leqq 0$. If $0 < \sigma < \delta$, then $\psi'(0) < 0$ and $\psi'(1) < 0$. There must be a first $\hat{x} \in (0, 1)$ such that $\psi(\hat{x}) = 0$ and $\psi'(\hat{x}) \geqq 0$. However, from (3.10), $\psi'(\hat{x}) = \sigma\hat{x} - \delta \leqq \sigma - \delta$, a contradiction. Therefore, $\delta \leqq \sigma$. If $\delta < 0$, then $\psi'(0) > 0$ and $\psi'(1) > 0$. At the first positive zero $\hat{x}$ of ψ we would have $\psi'(\hat{x}) \leqq 0$. However, $\psi'(\hat{x}) = \sigma\hat{x} - \delta > 0$, a contradiction. Therefore, it must be the case that $0 \leqq \delta \leqq \sigma$.

Next, suppose that $\alpha < -c$. Then, $s'(0) > 0$ and $s'(1) > 0$. At the next zero $\hat{x}$ of s we would have $s(\hat{x}) \geqq 0$ and $s'(\hat{x}) \leqq 0$. However, from (3.9) $s'(\hat{x}) = \psi^2(\hat{x}) - (\alpha + c) > 0$, a contradiction. Therefore, $\alpha \geqq -c$. We assume for the sake of contradiction that there is an unbounded, increasing sequence $\{\sigma_i\}_{i \geqq 0}$ such that for each i there is an $\alpha_i \in [-c, 0]$ and $\delta_i \in [0, \sigma_i]$ for which (3.9)-(3.12) has a solution. For notational simplicity, we omit the subscripts.

Let $L \equiv \tan(1)$ and define $M = 2(L^2 + 32L + 1)^{1/2}$ and

$$\sigma^* = \max\left\{256 \cdot M + 16 \cdot M \cdot L, \frac{112}{7} M + 14 \cdot M \cdot L\right\}.$$

We assume that $\sigma > \sigma^*$ and that there is an $\alpha \in [-c, 0]$, $\delta \in [0, \sigma]$ such that (3.9)-(3.12) has a solution. There are two subcases to consider.

Case (i). $0 \leqq \delta \leqq \sigma/2$. Consider the intervals $I_1 = [\frac{5}{8}, \frac{3}{4}]$ and $I_2 = [\frac{3}{4}, \frac{7}{8}]$. Suppose that $\psi(\frac{7}{8}) < -M$. If $\psi \leqq -M$ on all of I_1 then $\psi^2 \geqq M^2$, and hence $s' \geqq M^2 - L^2 - 1$ on I_1. Thus, $s(x) \geqq (M^2 - L^2 - 1)(x - \frac{5}{8}) - L$ and it follows that $s(\frac{3}{4}) \geqq (M^2 - L^2 - 1)/8 - L > L$, a contradiction. Therefore, $\psi(\hat{x}) > -M$ at some $\hat{x} \in [\frac{5}{8}, \frac{3}{4}]$. If $\psi(\tilde{x}) = -M$ at some first $\tilde{x} \in (\hat{x}, \frac{7}{8}]$ then $\psi'(\tilde{x}) \leqq 0$. However, $\psi'(\tilde{x}) \geqq \sigma(\tilde{x} - \frac{1}{2}) + 2sM > -2LM + \sigma/8 > 0$, a contradiction. Therefore, $\psi \geqq -M$ on all of I_2. Next, suppose that $\psi(\frac{3}{4}) > M$. Then it follows as above that $\psi > M$ on all of I_2. But then $s' \geqq M^2 - L^2 - 1$ on I_2 and an integration of this inequality leads to $s(x) \geqq (M^2 - L^2 - 1)(x - \frac{3}{4}) - L$. Hence, $s(\frac{7}{8}) > L$, a contradiction. Therefore, $\psi(\frac{3}{4}) < M$. Suppose that $\psi(\hat{x}) = M$ at some first $\hat{x} \in (\frac{3}{4}, \frac{13}{16})$. Then $\psi'(\hat{x}) > 0$. It then follows as above that $\psi > M$ for all $x \in (\frac{13}{16}, \frac{7}{8})$. Again, from (3.9) it follows that $s' \geqq M^2 - L^2 - 1$ on $(\frac{13}{16}, \frac{7}{8})$. Integrating, we obtain $s(x) \geqq (M^2 - L^2 - 1)(x - \frac{13}{16}) - L$. Therefore, $s(\frac{7}{8}) \geqq (M^2 - L^2 - 1)/16 - L > L$, a contradiction. We conclude that $\psi \leqq M$ for all $x \in (\frac{3}{4}, \frac{13}{16}) \equiv I_3$. On I_3, $\psi' \geqq \sigma/4 - 2LM$, so that $\psi(x) \geqq (\sigma/4 - 2LM)(x - \frac{3}{4}) - M$. Thus, $\psi(\frac{13}{16}) \geqq (\sigma/4 - 2LM)/16 - M > M$, a contradiction. We have shown that if $0 \leqq \delta \leqq \sigma/2$, there can be no solution to (3.9)-(3.12).

Case (ii). $\sigma/2 \leqq \delta \leqq \sigma$. Consider the intervals, $J_1 = [\frac{1}{8}, \frac{1}{4}]$ and $J_2 = [\frac{1}{4}, \frac{3}{8}]$. Suppose that $\psi(\hat{x}) < -M$ at some $\hat{x} \in [\frac{1}{8}, \frac{1}{4}]$. If $\psi(\tilde{x}) = M$ at some first $\tilde{x} \in (\hat{x}, \frac{3}{8}]$, then $\psi'(\tilde{x}) \geqq 0$. However, $\psi'(\tilde{x}) = \sigma(\tilde{x} - \frac{1}{2}) + 2sM$, so that $\psi'(\tilde{x}) < -\sigma/8 + 2LM < 0$, a contradiction.

Therefore, $\psi < -M$ on J_2. Thus, $s' \geqq M^2 - L^2 - 1$ on J_2 and it follows that $s(x) \geqq -L + (M^2 - L^2 - 1)(x - \frac{1}{4})$ on J_2 with $s(\frac{3}{8}) > L$, a contradiction. Thus, it must be the case that $\psi > -M$ on all of $J_1 = [\frac{1}{8}, \frac{1}{4}]$. If $\psi(\hat{x}) < M$ at some $\hat{x} \in [\frac{1}{4}, \frac{3}{8}]$, then it follows as above that $\psi < M$ on all of $J_1 \subseteq [\frac{1}{8}, \hat{x}]$. That is, $|\psi| < M$ for $\frac{1}{8} \leqq x \leqq \frac{3}{8}$. However, it follows from (3.10) that $\psi' \leqq -\sigma/4 + 2LM$ for all $x \in J_1$, hence $\psi(x) \leqq (2LM - \sigma/4)/8 + M < -M$, a contradiction. Therefore, it must be the case that $\psi > M$ for all $x \in J_2$. Thus, $s' \geqq M^2 - L^2 - 1$ and $s \geqq -L + (M^2 - L^2 - 1)(x - \frac{1}{4})$ on J_2. But, then, $u(\frac{3}{8}) \geqq (M^2 - L^2 - 1)/8 - L > L$, a contradiction. This eliminates the possibility of solutions to (3.9)-(3.12) for $\sigma/2 \leqq \delta \leqq \sigma$.

These two cases show that if σ is sufficiently large, there can be no solutions to (3.9)-(3.12) with $c \leqq 1$. □

Remark. The theorem requires that $c \leqq 1$. The reason for this is that we must bound $u(x)$ and this bound depends on $\tan(c)$. Thus, we could improve this bound slightly, by requiring that $c < \pi/2$. As we noted in our previous paper [1], if the size of the attraction to the limit cycle Λ is large compared to the diffusion d (i.e., c is large), then this phenomenon will not occur. Instead, as the gradient in frequency σ increases, the phase-locked periodic solutions to (3.1)-(3.3) will break up into complex high-dimensional tori. In fact, we can obtain bounds on the largest allowable frequency gradient σ as a function of the ratio d/Λ, before phase-locking is lost (in preparation). Our results here and in the previous paper show that if the diffusivity is large, phase-locking persists as σ increases until the amplitude of the solution goes to zero. The rest state remains stable for all higher values of σ.

The results of this theorem are essentially the continuum analogue of the results in [4] for a pair of coupled λ-ω oscillators. This may lead us to generalize the stability results of § 2 to an analogous continuum model

$$\frac{\partial u}{\partial t} = (1 + \sigma x) F(u) + d \frac{\partial^2 u}{\partial x^2} \tag{3.13}$$

subject to the boundary conditions

$$\frac{\partial u}{\partial x}(1, t) = \frac{\partial u}{\partial x}(0, t) = 0 \tag{3.14}$$

where d and σ are scalars. We have numerically solved (3.13)-(3.14) using the Wilson-Cowan equations as described in § 2. We find that as σ increases the amplitude of the oscillation tends to zero and for a large enough value of σ, the rest state appears to become stable. If we linearize (3.13) about the equilibrium state and let A denote the Jacobian of F at this equilibrium, we obtain

$$\frac{\partial u}{\partial t} = (1 + \sigma x) A u + d \frac{\partial^2 u}{\partial x^2} \tag{3.15}$$

along with the boundary conditions (3.14). As in § 2, we let ν and ϕ be an eigenvalue-eigenvector pair for the matrix A, i.e., $A\phi = \nu\phi$. We suppose that $u(x, t) = \phi z(x)\, e^{\lambda t}$, where $z(x)$ is a complex scalar function of x, and λ is a complex parameter. It is seen that z satisfies

$$\lambda z = (1 + \sigma x)\nu z + d z'', \tag{3.16}$$

$$z'(0) = z'(1) = 0. \tag{3.17}$$

This is similar to the problem solved by Theorem 3.1. Based on this, we conjecture that if $\mathrm{Re}(\nu)/\mathrm{Im}(\nu)$ is sufficiently small (as was required in § 2), then for d and σ

large enough, the eigenvalues λ of (3.16)-(3.17) have negative real parts. The proof of this is not as simple as that of Theorem 3.1, and it is hoped that we can eventually sharpen the statement and prove it.

For two-dimensional problems, we can intuitively understand what is going on from a geometric point of view as follows. Think of the equal strength diffusion as trying to pull points on each of the spatially distributed limit cycles together. For strong frequency gradients, the points (if uncoupled) would not be close to each other as time increases. Thus, nearby spatial points that are far away in phase will be pulled toward each other by the strong diffusion. Since they are at distant phases, the "easiest" way to pull them closer in state is through the equilibrium point about which they oscillate. Thus, the limit cycles will be shrunken to a single point by the strong diffusive forces (see Fig. 2).

4. Uniqueness of the trivial solution. We turn now to the question of uniqueness of the zero rest state $(u, v) = (0, 0)$ as a solution to (3.1)-(3.3). We consider the case $q = 0$, the "no twist" situation for which the uncoupled limit cycle of (3.1)-(3.3) has "radial isochrons" (see [3] for a discussion of this assumption and its consequences for two coupled oscillators). We look for periodic solutions of (3.1)-(3.3) that are of the form:

$$(u(x, t), v(x, t)) = (r(x) \cos(\theta(x, t)), r(x) \sin(\theta(x, t))), \tag{4.1}$$

$$\theta(x, t) = \int^x \phi(s)\, ds + \left(\Omega + \frac{\sigma}{2}\right) t. \tag{4.2}$$

We see that if a solution of this form exists, then it is t-periodic with period $2\pi/(\Omega + \sigma/2)$. As in § 3, we introduce the new variable $s(x) = r'(x)/r(x)$, so that in these variables, (3.1)-(3.3) become:

$$\begin{aligned}
&\text{(a)} \quad r' = rs, \\
&\text{(b)} \quad s' = \alpha^2(r^2 - 1) + \phi^2 - s^2, \\
&\text{(c)} \quad \phi' = \alpha^2\sigma(x - \tfrac{1}{2}) - 2s\phi, \\
&\text{(d)} \quad s(0) = \phi(0) = 0, \qquad 0 \leqq r(0) \leqq 1, \\
&\text{(e)} \quad s(1) = 0
\end{aligned} \tag{4.3}$$

where $\alpha^2 = 1/d$. In [1], we have shown that for each $r(0) \in [0, 1]$ and $\alpha^2 < 1$, there was a σ such that (4.3) has a solution. We will now show that for σ sufficiently large, the only solution to (4.3) is $r \equiv 0$, corresponding to $(u, v) = (0, 0)$.

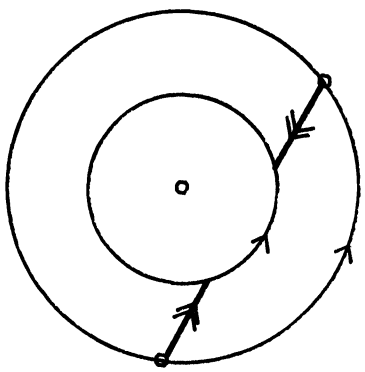


FIG. 2. *Two points that initially are close in phase are pulled apart by differences in frequency and the resultant limit cycle is "pinched" together by strong diffusive forces.*

THEOREM 4.1. *Let* $0<\alpha\leqq 1$ *and* $q=0$. *Then there is a* $\sigma^*=\sigma^*(\alpha)>0$ *such that if* $\sigma>\sigma^*$ *then there is no value of* $r(0)\geqq 0$ *for which* (4.3) *has a solution.*

As the proof is again rather technical, we will sketch an outline. The first step is to set $L=\alpha\tan(\alpha)$, $0<\alpha<1$ and prove that if $s(x_0)=L$ for some $x_0\in(0,1)$, then $s(1)\neq 0$. Thus, a solution to (4.3) must satisfy $s(x)<L$ for all $x\in(0,1)$. We assume, for the sake of contradiction, that a solution of the boundary value problem does indeed exist, for some $\alpha\in(0,1)$ and some $r(0)\in[0,1)$. To obtain a contradiction, we need to analyze the equation for u. Since ψ appears in the equation, we must first obtain lower bound estimates on ψ^2 by analyzing (4.3)(c). Thus, the next step of the proof is to prove that there are values $\delta\in(0,\frac{1}{4})$ and $C>0$ independent of σ such that $\psi^2\geqq C\sigma^2$ for all $x\in[\delta/2,\delta]$. With this estimate in hand we then turn to the s equation. First we obtain a lower bound on s, namely, $-\tan(\sigma)<s<L$ for $x\in[\delta/2,\delta]$. It then follows that $s^2\leqq Q^2$ on $[\delta/2,\delta]$, where $Q=\max\{L,\tan(\delta)\}$. These estimates on s^2 and ψ^2 are then substituted into (4.3)(b), and we obtain $s'\geqq-(\alpha^2+Q^2)+C\sigma^2$ for $x\in[\delta/2,\delta]$. This last inequality is used to show that for large σ, s must reach L at some $x_0\in[\delta/2,\delta]$. As shown earlier, this implies that $s(1)\neq 0$, and we have the needed contradiction.

Proof of Theorem 4.1. We first derive a criterion which guarantees that (4.3)(e) cannot hold. From (4.3)(b) it is seen that

$$s'\geqq-(\alpha^2+s^2).$$

Integrating this inequality leads to

$$\frac{1}{\alpha}\tan^{-1}(s(x)/\alpha)\geqq\frac{1}{\alpha}\tan^{-1}(s(x_0)/\alpha)-(x-x_0)\quad\text{for all } x_0.$$

Letting $x_0\in(0,1)$ and $x=1$, we find

$$\frac{1}{\alpha}\tan^{-1}(s(1)/\alpha)\geqq\frac{1}{\alpha}\tan^{-1}(s(x_0)/\alpha)-1+x_0.$$

Thus, if the right-hand side is positive, i.e.,

$$\frac{1}{\alpha}\tan^{-1}(s(x_0)/\alpha)>1-x_0,$$

then the left-hand side is also positive and $s(1)\neq 0$. It suffices to prove that there is a $\sigma^*>0$ such that if $\sigma>\sigma^*$, there is an $x_0(\sigma)\in(0,1)$ and a $s(x_0)>0$ with $\tan^{-1}(s(x_0)/\alpha)=\alpha$, i.e.,

$$s(x_0)=\alpha\tan(\alpha)\equiv L.$$

If $0<\alpha<\pi/2$, L is finite and positive; furthermore, L is independent of σ. We will show that as σ increases, s rises to L at some $x_0\in(0,1)$. For technical reasons, we require that x_0 lies in $(0,\frac{1}{2})$.

Suppose that for some $r(0)\in[0,1)$ and all large σ that $s(x)<L$ for all $x\in(0,\frac{1}{4})$. We note that $\phi(x)<0$ for all $x\in(0,\frac{1}{4}]$. For otherwise, since $\phi(0)=0$ and $\phi'(0)=-\alpha^2\sigma/2<0$, there must be some first $\hat{x}\in(0,\frac{1}{4}]$ for which $\phi(\hat{x})=0$ and $\phi'(\hat{x})\geqq 0$. But, from (4.3)(c) $\phi'(\hat{x})=\alpha^2\sigma(\hat{x}-\frac{1}{2})<0$, a contradiction. Thus if $s(x)<L$ for all $x\in(0,\frac{1}{4})$, then from (4.3)(c),

$$\phi'<\alpha^2\sigma(x-\tfrac{1}{2})-2L\phi.$$

Integrating this inequality leads to

$$\phi(x)<\frac{\alpha^2\sigma}{2L}\left[x-\frac{1}{2L}-\frac{1}{2}+\left(\frac{1}{2}+\frac{1}{2L}\right)e^{-2Lx}\right]\equiv\frac{\alpha^2\sigma}{2Lg(x)}.$$

The function $g(x)$ satisfies the following:

$$g(0)=0, \quad g'(0)<0, \quad g''(0)>0 \quad \text{for all } x>0.$$

Thus, $g(x)<0$ for x small and positive, but for x sufficiently large, $g(x)>0$, since $g''>0$. Let $\tilde{x}=\ln(1+L)/2L$ denote the value of x at which g' vanishes and let $\delta=\min\{\frac{1}{4}, \tilde{x}/2\}$. Note that δ is independent of σ. Now consider the interval $[0, \delta]$. $g'(x)<g'(\delta)<0$ for all $x\in[0,\delta]$. Let $-M=g'(\delta)$, so that $g'(x)<-M$ for $x\in[0,\delta]$. Thus, $g(x)<-Mx$ and $\phi(x)<-\alpha^2\sigma Mx/2L$ for $x\in[0,\delta]$. In particular, over the interval $I\equiv[\delta/2, \delta]$, we have $\phi(x)<-\alpha^2\sigma M\delta/4L$. So,

$$\phi^2(x)\geqq\sigma^2 C, \qquad x\in I \tag{4.4}$$

where $C=\alpha^4M^2\delta^2/16L^2$ is independent of σ.

We turn to the s equation. We have assumed that $s\leqq L$ in I. We compute a lower bound on s. Since $\alpha\leqq 1$, $s'\geqq-(1+s^2)$ from (4.3)(b) for all $x\geqq 0$ as long as a solution exists. Integrating this over $[0, x)$, we find that $s(x)\geqq-\tan(x)$ for all $x\in[0,\frac{1}{2}]$. This implies that $s(x)\geqq-\tan(\delta)$ for $x\in I$ and therefore

$$-\tan(\delta)\leqq s(x)\leqq L \quad \text{for } x\in I. \tag{4.5}$$

Let $Q=\max\{L, \tan(\delta)\}$. Then $s^2\leqq Q^2$ in I and Q depends only on α. Equations (4.3)(b), (4.4), and (4.5) imply that

$$s'\geqq-(\alpha^2+Q^2)+C\sigma^2 \quad \text{for } x\in I.$$

We integrate this from $\delta/2$ to δ and obtain

$$s(x)\geqq s(\delta/2)+[C\sigma^2-(\alpha^2+Q^2)](x-\delta/2).$$

Since $s(\delta/2)\geqq-\tan(\delta)$, we find

$$s(\delta)\geqq-\tan(\delta)+[C\sigma^2-(\alpha^2+Q^2)]\delta/2.$$

C and Q are independent of σ, so there is a $\sigma^*=\sigma^*(\alpha)$, independent of $r(0)$ such that for $\sigma>\sigma^*$, $s(\delta)>L$. This implies that $s(x_0)=L$ for some $x_0\in(0,\delta)$, contradicting our assumption that $s(x)<L$ for all $x\in(0,1)$. Thus, $s(1)\neq 0$. □

Remarks. (1) We emphasize that this theorem and the previous theorem depend on the fact that the diffusion d was sufficiently large. In fact, if d is too small, this phenomenon does not happen; rather, desynchronized solutions appear. Thus, the loss of the periodic solution to a stable equilibrium depends on several criteria: the frequency gradient must be large, the diffusion must be sufficiently strong, and finally, the diffusion must be close to scalar. The last condition seems to be needed through the heuristic argument given at the end of § 3 and Fig. 2. In another paper (in preparation) the case of very weak diffusion along with $O(1)$ frequency gradients is explored for the special system (3.1)-(3.3) as well as for general reaction-diffusion equations.

(2) Numerical integration of (3.1)-(3.3) indicates that with strong diffusion, the periodic solution shrinks to zero even in the presence of the twist term q. This term, of course, has no effect on the stability of the rest state, so we believe the uniqueness also holds for $q\neq 0$.

REFERENCES

[1] G. B. ERMENTROUT AND W. C. TROY, *Phaselocking in a reaction-diffusion system with a linear frequency gradient*, SIAM J. Appl. Math., 46 (1986), pp. 359-367.

[2] K. BAR-ELI, *On the stability of coupled chemical oscillators*, Phys. D., 14 (1985), pp. 242-252.

[3] D. ARONSON, G. B. ERMENTROUT, AND N. KOPELL, *Amplitude response of coupled oscillators*, Phys. D (1989), to appear.

[4] G. B. ERMENTROUT AND N. KOPELL, *Oscillator death in systems of coupled neutral oscillators*, SIAM J. Appl. Math., to appear.

[5] S. SMALE, *A mathematical model of two cells coupled via Turing's equation*, in Lectures on Mathematics in the Life Sciences 6, J. D. Cowan, ed., American Mathematical Society, Providence, RI, 1974, pp. 17-26.

[6] G. B. ERMENTROUT, *Stable small-amplitude solutions in reaction-diffusion systems*, Quart. Appl. Math., April (1981), pp. 61-86.

[7] H. R. WILSON AND J. D. COWAN, *Excitatory and inhibitory interactions in localized populations of model neurons*, Biophys. J., 12 (1973), pp. 1-24.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1447-1458, November 1989

010

THE UTILITY OF AN INVARIANT MANIFOLD DESCRIPTION OF THE EVOLUTION OF A DYNAMICAL SYSTEM*

A. J. ROBERTS†

Abstract. The long-time evolution of a physical system may be dominated by a small number of modes. An invariant manifold description of the asymptotic evolution, briefly discussed here in two simple examples, can give a powerful and useful view of the physical system. The relationship between invariant manifolds and centre manifolds is also discussed, as is the relationship between invariant manifolds and modal numerical approximations. Many classical approximations in fluid mechanics can be viewed as approximating the evolution on an approximate invariant manifold.

Key words. asymptotic expansions, evolution equations, invariant manifolds, centre manifold theory, numerical approximation

AMS(MOS) subject classifications. 34E05, 35K22, 58F40, 76A99

1. Introduction. Centre manifold theory can provide a powerful asymptotic description of the long-time evolution of a physical system. For some simple examples see Carr [5] or Guckenheimer and Holmes [8]; for more sophisticated applications see Roberts [15]-[18], Bernhoff [3], Arneodo, Coullet, and Spiegel [1], Arneodo and Thoul [2], Coullet and Spiegel [7], and Mercer and Roberts [9]. In essence, centre manifold theory describes the evolution of a system in terms of the evolution of a relatively small number of dominant modes. These modes are those that have marginal stability (the linear growth rates are essentially zero) and are therefore long-lived. However, in some practical applications the requirement that the retained modes are marginal is unduly restrictive. Here we investigate two simple dynamical systems and derive a method to describe their long-time asymptotic behavior when the decay rates of their dominant modes are not "essentially zero."

That this sort of analysis will be of relevance can be seen by considering some examples; the theory of quasistationary probability distributions (see Parson and Pollett [12] or Pollett and Roberts [13]) furnishes one. A system with a given number of states and given probabilistic transition rates typically has an absorbing state, corresponding to a zero eigenvalue, to which the system eventually evolves. However, it is sometimes the case that a long-lived transient state (i.e., quasistationary), corresponding to a small negative eigenvalue, is of vital importance. To be effective, the asymptotic description of the system's evolution must include both the absorbing state and the nonmarginal quasistationary state. As another example, consider any infinite-dimensional system with a continuous spectrum (see Bernhoff [3] for an example in convection, or Roberts [16] for a model system). Somewhere in such a continuous spectrum, centre manifold theory requires a dividing line to be drawn between those modes whose linear growth and decay rates are essentially zero and those modes whose decay rates are definitely nonzero. Such a dividing line in a continuum is clearly more or less arbitrary and is an uncomfortable thing to draw. Some rationale for including or excluding modes corresponding to a continuous spectrum is needed. As a last example, Taylor [25] showed that longitudinal dispersion in channels and rivers may be greatly enhanced by the effects of shear in the current. However, the asymptotic equation he derived

* Received by the editors November 11, 1987; accepted for publication (in revised form) January 20, 1989.
† Department of Applied Mathematics, University of Adelaide, G.P.O. Box 498, Adelaide, South Australia.

appears to be of limited practical value, as it takes too long for the system to evolve onto the centre manifold implicit in his asymptotic equations. If, instead of just one dominant marginal mode, one or two additional nonmarginal modes are retained (see Chikwendu and Ojiakor [6], Smith [24], or Roberts [21]), then the resultant asymptotic description promises to be much more useful in its application.

The task of this paper is to describe the modifications, and their implications, of the formal centre manifold procedure described by Coullet and Spiegel [7] and Roberts [15]-[18], when the dominant modes do not necessarily have an essentially zero growth rate. The result is the description of an invariant manifold of the system (rather than a centre manifold characterised by eigenvalues having precisely zero growth rate). This is introduced in § 2, where a simple dynamical system and its evolution are discussed. The occurence of zero divisors in the resultant asymptotic formula has implications for both the range of validity of this invariant manifold description, discussed in § 3, and for the typically divergent character of the centre manifold description of a bifurcation problem, discussed in § 4.

A general, formal procedure for deriving such descriptions is then introduced in the context of a particular problem, that of finding a quasistationary probability distribution. The central approximation, which involves only elementary algebraic concepts, is discussed in § 5. Then in § 6 the method of Coullet and Spiegel [7], appropriate when the system is subject to small nonlinearities or other perturbations, is modified to give a description of an invariant manifold (rather than a centre manifold). It is this formal procedure that is expected to be of immense practical utility.

Furthermore, this invariant manifold viewpoint provides a rational basis for the derivation of appropriate initial conditions for such low-dimensional descriptions of the evolution of the dynamical system. This is developed in a subsequent paper (Roberts [20]).

In § 7, it is argued that truncated-modal numerical solutions of a set of equations can be viewed as a particularly simple approximation of an invariant manifold. This view then indicates how such numerical models may be improved without increasing the number of retained modes. It also suggests how to economically solve stiff sets of ordinary differential equations (see Roberts [19]).

Last, § 8 discusses how the concept of an invariant manifold, as a state which the system will ultimately approach, may be generalised. The concept is useful in physical situations where the invariant manifold is not necessarily approached at large times, but is in some sense central to the behaviour of the physical system. Some applications of this are to the irrotational approximation and vortex dynamics, pole solutions of solitons, quasigeostrophy, incompressibility, and the continuum theory of monatomic gases.

2. A formal invariant manifold description. Here a formal procedure to calculate an invariant manifold of a simple example system is given. The results are then used to illustrate what appear to be general features of such an invariant manifold description. Consider the nonlinear autonomous pair of different equations:

$$\dot{x} = \lambda x - xy, \tag{2.1a}$$

$$\dot{y} = -y + x^2, \tag{2.1b}$$

where $(\dot{})$ denotes d/dt and λ is some parameter. This is a system that is often used to illustrate the pitchfork bifurcation as it has fixed points at $(0, 0)$ and $(\pm\sqrt{\lambda}, \lambda)$. The trajectories are illustrated in Fig. 1 of Roberts [14], and a centre manifold analysis, based on λ being small (i.e., essentially zero) is given in § 3 of Roberts [15]. The

problem we now address is: what is the long-time behaviour of (2.1) if we consider λ to be small but not necessarily near zero?

Suppose $|\lambda|$ is significantly smaller than 1 but is not so small that λx can be considered a "nonlinear" term (see Carr [5]). Linearly, (2.1) has trajectories

$$y \sim A|x|^{-1/\lambda} \quad \text{as } |x| \to 0, \tag{2.2}$$

which, for small λ, are quite "flat." Hence, to an error that decreases exponentially with time (as e^{-t} since the other eigenvalue of the linearised system is -1) and to an error of order $|x|^{-1/\lambda}$, we could describe the long-time evolution of the linearised version of (2.1) as

$$\dot{x} \approx \lambda x \quad \text{on } y \approx 0. \tag{2.3}$$

This is just an elementary description of the stability of the fixed point at the origin. To produce an improved asymptotic description of (2.1) we need to include the effects of the nonlinearity. Suppose that $y = h(x; \lambda)$ is an invariant manifold of (2.1) on which the system evolves, according to $\dot{x} = \lambda x - xh(x; \lambda)$ by (2.1a). Then

$$\dot{y} = h'\dot{x} = h'(\lambda x - xh) \quad \text{and} \quad \dot{y} = -h + x^2,$$

where a prime sign denotes $\partial/\partial x$ keeping λ constant, and so the manifold $y = h(x; \lambda)$ must satisfy the differential equation

$$h + \lambda x h' = x^2 + xhh'. \tag{2.4}$$

By virtue of the linear result (2.3) we recognise that we should find a manifold $y = h(x; \lambda)$ that is "flat" near the origin; that is, (2.4) should be solved such that

$$h = O(x^2) \quad \text{as } |x| \to 0. \tag{2.5}$$

Typically, equations such as (2.4) are solved via an asymptotic expansion (for an exception see § 2 in Roberts [15]), which is what we now do. Suppose h may be expanded in the formal power series

$$h \sim \sum_{n=2}^{\infty} h_n x^n \quad \text{as } x \to 0 \tag{2.6}$$

(actually a finite truncation of this sum may be all that is valid; see the discussion in the next section). Upon substituting this into (2.4) and grouping all terms of the same power of x, we find

$$(1 + n\lambda)h_n = \begin{cases} 1, & n = 2, \\ \sum_{m=2}^{n-2} m h_m h_{n-m}, & n = 3, 4, \cdots. \end{cases} \tag{2.7}$$

Solving these equations in succession, we easily deduce

$$h_2 = \frac{1}{(1+2\lambda)}, \quad h_4 = \frac{2}{(1+2\lambda)^2(1+4\lambda)}, \quad h_6 = \frac{12}{(1+2\lambda)^3(1+4\lambda)(1+6\lambda)}, \tag{2.8}$$

and so on ($h_n = 0$ for odd n by the symmetry of (2.1)). Thus the invariant manifold description we obtain is that system (2.1) settles exponentially quickly onto a manifold given by

$$y \approx \frac{x^2}{(1+2\lambda)} + \frac{2x^4}{(1+2\lambda)^2(1+4\lambda)} + \frac{12x^6}{(1+2\lambda)^3(1+4\lambda)(1+6\lambda)}, \tag{2.9}$$

on which its evolution is governed by

$$\dot{x} \approx \lambda x - \frac{x^3}{(1+2\lambda)} - \frac{2x^5}{(1+2\lambda)^2(1+4\lambda)} - \frac{12x^7}{(1+2\lambda)^3(1+4\lambda)(1+6\lambda)}. \tag{2.10}$$

3. Range of validity. The new features in the asymptotic description above arise from the form of the left-hand side of (2.4). In a centre manifold analysis λ is assumed to be asymptotically zero and the term $\lambda x h'$ would appear on the right-hand side as a perturbation term. However, here this term is of the same order as h itself and must appear on the left. Its effect is to cause divisors of the form $(1+n\lambda)$, and higher powers, to appear in the asymptotic description (see (2.8)-(2.10)). For negative λ, in particular for $-1 \leqq \lambda < 0$, this gives rise to the embarrassing possibility of dividing by zero. Furthermore, even if a division by precisely zero is avoided, a division by nearly zero, especially when compounded at higher order, is nearly as bad. Fortunately, this problem of zero divisors serves as a useful and informative warning.

Suppose that we have described an invariant manifold by a particular solution of (2.4), $y = h_p(x;\lambda)$. Other invariant manifolds are given by neighbouring solutions of (2.4), namely, $y = h_p + \hat{h}(x;\lambda)$, where $\hat{h}$ is small and satisfies the linearised equation

$$\hat{h} + \lambda x \hat{h} = x h_p' \hat{h} + x h_p \hat{h}'. \tag{3.1}$$

For small x the right-hand side of (3.1) is negligible, and so (as in (2.2)), $\hat{h} \sim A|x|^{-1/\lambda}$ as $x \to 0$ for arbitrary A. Hence, the invariant manifold described is unique only up to order $|x|^{-1/\lambda}$, but this is precisely the order at which a zero divisor occurs. Thus the zero divisor is indicative of the nonuniqueness of the invariant manifold. Consequently the invariant manifold is blurred by an amount $O(|x|^{-1/\lambda})$, and we conclude that there is little point in calculating terms of higher order than $|x|^{-1/\lambda}$ (it is at this order that the initial conditions determine which particular manifold to choose).

Indeed, the conclusion above is related to the property (Shub [22, p. 61]) that for $-1 \leqq \lambda < 0$ the invariant manifold is only guaranteed to be C^r where $r > -1/\lambda$. That is, the invariant manifold is typically not analytic, but it is differentiable to a finite order; thus the assumed asymptotic series (2.6) must be truncated before this order. (For $\lambda < -1$ no zero divisors appear, and this particular invariant manifold is C^∞. However, it is then of little interest to the long-term evolution of the dynamical system, as the decay on this invariant manifold is much quicker than the decay elsewhere.)

The above argument only applies for negative λ. For positive λ the invariant manifold is C^∞ and no small divisors appear. However, it is only C^∞ sufficiently close to the origin. At the finite amplitude fixed point $(\pm\sqrt{\lambda}, \lambda)$ the decay rates of the linearised equations are the slow $-\sqrt{\lambda}$ and the rapid -1; consequently we expect the invariant manifold that passes through the origin to be C^r at the finite-amplitude fixed point where $r < 1/\sqrt{\lambda}$. As discussed above, the invariant manifold relevant to the long-term evolution is thus "blurred" by an amount $O(A(x \pm \sqrt{\lambda})^{1/\sqrt{\lambda}})$, but such blurring of the invariant manifold is not readily apparent in the asymptotic description (2.8)-(2.10). Nonetheless, for such positive λ it may be appropriate to calculate terms of higher order than $|x|^{1/\sqrt{\lambda}}$. This is not so much to fix the whole of the invariant manifold, as to refine predictions about the location and nature of the finite-amplitude fixed points (or attractors).

4. Relation to the centre manifold analysis. A centre manifold analysis of system (2.1) has been described in § 3 of Roberts [15], and is based on the assumption that λ, x, and y are all asymptotically small. The results obtained in equation (3.49) of [15] are precisely the same as the small λ asymptotic expansion of the invariant manifold result obtained here in (2.9) and (2.10). For example, if the manifold $h(x;\lambda)$ is expanded in powers of λ, we find

$$y = h(x;\lambda) \sim (1 - 2\lambda + 4\lambda^2)x^2 + 2(1 - 8\lambda + 44\lambda^2)x^4 \quad \text{as } x, \lambda \to 0,$$

which is precisely equation (3.49c) of Roberts [15]. The first consequence of this observation is that this invariant manifold approach gives equivalent results, to any power in λ, as the centre manifold approach to this same problem.

The second consequence is that the comments on the range of validity of the invariant manifold results given in the previous section also apply to the centre manifold analysis. Thus, for λ too far from its critical value of zero, there is little use in endeavouring to be too precise about the centre manifold and the evolution on it.

Last, as noted by Sijbrand [23], we can observe that a centre manifold description of bifurcation will typically be asymptotic; only rarely will the resulting series converge. Expressions (2.8) for the coefficients h_n of the invariant manifold contain the divisors $(1+m\lambda)$. In fact, the coefficient h_{2n} will typically be of the form

$$h_{2n} = H_n \prod_m (1+2m\lambda)^{-[n/m]},$$

where H_n is some number independent of λ, and where [ ] denotes the integer part function. Thus the centre manifold results, which are equivalent to expanding h_n in powers of λ, will involve the expansions in λ of $(1+m\lambda)^{-p}$ for arbitrarily large m and will therefore have a zero radius of convergence. That is, a centre manifold expansion of bifurcation problems is typically divergent (asymptotic descriptions of invariant manifolds in bifurcation problems may often converge to some extent; see Roberts [19]). However, for problems that do not involve a bifurcation, a centre manifold description can converge. Three examples of this are described by Mercer and Roberts [9] for shear dispersion in a channel, Roberts [17] for slowly varying waves, and Pollett and Roberts [13] for quasistationary probability distributions.

5. A quasistationary probability distribution. Given that the concept of an invariant manifold is useful in the description of the long-time evolution of a system, we need to find a general method for deriving such an asymptotic description. The method will be a modification of that given by Coullet and Spiegel [7] (also explained and extended in Roberts [15]-[19]) to systems where the retained modes do not necessarily have an essentially zero growth rate. This section introduces the example problem of finding a quasistationary probability distribution, modified in the next section, where the general method of finding an invariant manifold and the evolution on it is discussed.

Consider a Markov process consisting of three states, labelled 1, 2, 3, where the probability vector $\mathbf{p}$ of the system evolves according to[1]

$$\frac{d\mathbf{p}}{dt} = Q\mathbf{p}, \tag{5.1}$$

in which the rate matrix Q is

$$Q = \begin{bmatrix} 0 & \lambda & \lambda \\ 0 & -\lambda-\frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & -\lambda-\frac{1}{2} \end{bmatrix}, \tag{5.2}$$

where λ is a positive parameter. Observe that since total probability must be conserved, the column sums of Q are all zero. The problem is then to determine an appropriate description of the long-time behaviour of (5.1).

[1] Probabilists, please note that Q and $\mathbf{p}$ are the transpose of what you are used to. This is because the example is meant to be illustrative and almost everyone else is accustomed to this form for ordinary differential equations.

Since (5.1) is a linear system, its analysis is particularly simple and is essentially just an elementary change of basis problem. However, it is the viewpoint developed here (rather than the analytical details) that is very valuable and that provides the basis for the powerful formal procedure discussed in the next section.

We start by finding that the eigenvalues and eigenvectors of the matrix Q are

Eigenvalue	Eigenvector	Left eigenvector
0	$\mathbf{e}_1=\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\hat{\mathbf{z}}_1=[1 \quad 1 \quad 1]$
$-\lambda$	$\mathbf{e}_2=\begin{bmatrix}-2\\1\\1\end{bmatrix}$	$\hat{\mathbf{z}}_2=[0 \quad 1 \quad 1]$
$-(1+\lambda)$	$\mathbf{e}_3=\begin{bmatrix}0\\1\\-1\end{bmatrix}$	$\hat{\mathbf{z}}_3=[0 \quad 1 \quad -1]$

Clearly, state 1 is an absorbing state and so the probability vector $\mathbf{p}$ ultimately evolves to be $\mathbf{e}_1$. However, if λ is a number significantly smaller than $(1+\lambda)$, then the time until absorption is relatively long and the system has time to evolve into a quasi-stationary state (involving the eigenvector corresponding to λ, namely, $\mathbf{e}_2$) and only then does it slowly evolve into the absorbing state. Thus a significant asymptotic description of (5.1) involves not only the absorbing state corresponding to the eigenvalue zero, but also involves the quasistationary state corresponding to the small (but nonzero) eigenvalue. Thus we need an invariant manifold analysis rather than a centre manifold analysis.

The analysis proceeds by posing the ansatz

$$\mathbf{p}(t)=\mathbf{v}(\mathbf{x}) \quad \text{such that} \quad \frac{d\mathbf{x}}{dt}=\mathbf{g}(\mathbf{x}), \tag{5.3}$$

where $\mathbf{x}=[x_1, x_2]^T$ are the probabilities x_1 of being in the absorbing state, and x_2 the quasistationary state, thus $\mathbf{x}$ evolves slowly. Substituting this into the governing equation (5.1), using the chain rule, and letting $\mathbf{v}_\mathbf{x}$ denote the 3×2 matrix $[\partial v_i/\partial x_j]$ we find that $\mathbf{v}$ and $\mathbf{g}$ must satisfy

$$Q\mathbf{v}=\mathbf{v}_\mathbf{x}\mathbf{g}. \tag{5.4}$$

As this is linear we can solve it exactly by supposing

$$\mathbf{v}=V\mathbf{x} \quad \text{and} \quad \mathbf{g}=G\mathbf{x}, \tag{5.5}$$

where V is a 3×2 matrix and G is a 2×2 matrix. Thus, since (5.4) must be true for all x, we deduce that V and G must satisfy

$$QV=VG. \tag{5.6}$$

This is a type of eigenproblem. Normally we would choose V to diagonalise Q (remember that we are interested only in the small eigenvalues of Q); that is, the columns of V are eigenvectors of Q, and G is the diagonal matrix of the corresponding

eigenvalues (the small ones only). In a centre manifold analysis these significant eigenvalues would be all zero; here they are 0 and $-\lambda$. This fails if Q is nondiagonalisable, in which case G may be chosen in Jordan form and (5.6) is then solved. However, the conventional practice of choosing G to be either diagonal or in Jordan form need not be followed. The choice of how to solve (5.6) is arbitrary (as long as the eigenvalues of G are the significant small ones); it all depends upon what the "amplitudes" $\mathbf{x}$ will represent in the asymptotic description. Here it is desirable (but not essential) for the evolution equation $d\mathbf{x}/dt = G\mathbf{x}$ itself to represent a Markov process, thus describing the evolution between two distinct states. Hence each row of V is chosen to have at most one nonzero element and the column sums of V are all 1 (that the column sums of G are zero then follows immediately).

Next, the invariant manifold of interest is the one spanned by the eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$, these being the long-lived modes as they correspond to the small eigenvalues zero and $-\lambda$. Imposing the above discussed constraints, we change the basis of this subspace to

$$\mathbf{q}_1 = \mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{q}_2 = \mathbf{e}_1 + \tfrac{1}{2}\mathbf{e}_2 = \begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \tag{5.7}$$

and deduce the solution

$$V = [\mathbf{q}_1 | \mathbf{q}_2] = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} \quad \text{and correspondingly} \quad G = \begin{bmatrix} 0 & \lambda \\ 0 & -\lambda \end{bmatrix}. \tag{5.8}$$

That is, the invariant manifold description of the evolution of (5.1) is that

$$\mathbf{p} = \mathbf{q}_1 x_1 + \mathbf{q}_2 x_2 \quad \text{such that} \quad \frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & \lambda \\ 0 & -\lambda \end{bmatrix} \mathbf{x}. \tag{5.9}$$

Thus, to an error that decreases as does $\exp[-(1+\lambda)t]$, we conclude that the long-time evolution of (5.1) is from the quasistationary state $\mathbf{q}_2$ to the absorbing state $\mathbf{q}_1$ at the rate λ.

6. The formal invariant manifold procedure. Problems of the type discussed in the previous section are elementary, but, being linear, they are fundamental to a perturbative analysis of nonlinear systems or of a family of similar systems. In this section a formal method is outlined for deriving the invariant manifold description of a system when the description is in the typically necessary form of an asymptotic series (as in § 2). This method is a modification of that given by Coullet and Spiegel [7], and to be more or less definite, the method is described in the context of a particular problem.

Suppose that (5.1) is modified in some manner by the inclusion of some asymptotically small linear or nonlinear terms $\mathbf{r}(\mathbf{p})$ to

$$\frac{d\mathbf{p}}{dt} = Q\mathbf{p} + \mathbf{r}(\mathbf{p}). \tag{6.1}$$

The procedure to find an asymptotic description of the corresponding invariant manifold (possibly curved) and the evolution on it is to assume that the unknowns $\mathbf{v}$ and $\mathbf{g}$ may be expanded in an asymptotic series. That is, we pose

$$\mathbf{p}(t) = \mathbf{v}(\mathbf{x}) \sim \sum_{k=0}^{\infty} \mathbf{v}^k(\mathbf{x}) \quad \text{such that} \quad \frac{d\mathbf{x}}{dt} = \mathbf{g}(\mathbf{x}) \sim \sum_{k=0}^{\infty} \mathbf{g}^k(\mathbf{x}), \tag{6.2}$$

where all terms of "order" k are grouped into $\mathbf{v}^k$ and $\mathbf{g}^k$ (the precise definition of order depends upon the form of the perturbing terms $\mathbf{r}(\mathbf{p})$ and need not concern us here). Upon substituting the ansatz (6.2) into (6.1), using the chain rule, and grouping all terms of the same order together (see Roberts [15], [16] for similar details), we find the sequence of equations

$$Q\mathbf{v}^k = \sum_{l=0}^{k} \mathbf{v}_{\mathbf{x}}^l \mathbf{g}^{k-l} - \mathbf{r}^k \tag{6.3}$$

to be solved, where $\mathbf{r}^k$ is a function of $\mathbf{v}^0, \cdots, \mathbf{v}^{k-1}$ that depends upon the form of the perturbation $\mathbf{r}(\mathbf{p})$, and where $\mathbf{r}^0 = 0$.

The $k=0$ equation is just the linear pseudo-eigenproblem (5.4) solved in the previous section. The relevant solution is $\mathbf{v}^0 = V\mathbf{x}$ and $\mathbf{g}^0 = G\mathbf{x}$, where V and G are given by (5.8).

The real interest of this section is in calculating the higher-order correction terms $\mathbf{v}^k$ and $\mathbf{g}^k$. At order k, (6.3) has the form

$$Q\mathbf{v}^k - \mathbf{v}_{\mathbf{x}}^k G\mathbf{x} = V\mathbf{g}^k + \mathbf{s}^k, \tag{6.4}$$

where $\mathbf{s}^k$ is known, as it depends only upon lower-order quantities that have already been calculated. This equation is to be solved for $\mathbf{v}^k$ and $\mathbf{g}^k$ (for clarity the superscript k will be omitted henceforth). It has to be solved independently of whatever values $\mathbf{x} = [x_1, x_2]^T$ may take, and so we consider x_1 and x_2 to be independent indeterminants (see Coullet and Spiegel [7]); that is, we consider (6.4) to be an equation that is a multinomial in x_1 and x_2. Thus we try for a solution in the form

$$\mathbf{v} = \sum_{m,n} \mathbf{v}_{mn} x_1^m x_2^n, \qquad \mathbf{g} = \sum_{m,n} \mathbf{g}_{mn} x_1^m x_2^n \quad \text{supposing } \mathbf{s} = \sum_{m,n} \mathbf{s}_{mn} x_1^m x_2^n, \tag{6.5}$$

where the sums may go from zero to some limit.

Upon substituting (6.5) into (6.4), equating like powers of $x_1^m x_2^n$, and using the particular form of G given in (5.8) we obtain

$$(Q + n\lambda I)\mathbf{v}_{mn} = \begin{cases} V\mathbf{g}_{mn} + \mathbf{s}_{mn}, & n = 0, \\ V\mathbf{g}_{mn} + \mathbf{s}_{mn} + (m+1)\lambda \mathbf{v}_{m+1,n-1}, & n \geqq 1. \end{cases} \tag{6.6}$$

Note that for a different form of G the coupling terms in (6.6), here $\mathbf{v}_{m+1,n-1}$, between distinct equations would be different. In particular, if G is diagonal then there are no coupling terms (and the solution is easier).

For whatever range of m and n is appropriate, (6.6) is solved such that the equations $m+n =$ constant are solved in the sequence $n = 0, 1, 2, \cdots$ (because of the particular nature of the coupling term). The $n=0$ equation is

$$Q\mathbf{v}_{m0} = V\mathbf{g}_{m0} + \mathbf{s}_{m0}. \tag{6.7}$$

Now Q has a zero eigenvalue and is therefore singular. To solve this equation the right-hand side has to be in the range of Q, that is, orthogonal to the left eigenvector $\hat{\mathbf{z}}_1$. This gives one linear equation to determine the two unknowns of $\mathbf{g}_{m0}$; the other necessary equations come from the precise definition of what the "amplitudes" x_1 and x_2 represent. Typically we want $\mathbf{v}_{mn}$ to have no component of the tangent space to the invariant manifold (that space spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$); that is, we want $\mathbf{v}_{mn}$ to be orthogonal to $\hat{\mathbf{z}}_1$ and $\hat{\mathbf{z}}_2$. (More generally, we would want $\mathbf{v}_{mn}$ to have definite components of $\mathbf{e}_1$ and $\mathbf{e}_2$, depending upon what x_1 and x_2 are defined to be.) Premultiplying (6.7) by $\hat{\mathbf{z}}_2$ we then find

$$-\lambda(\hat{\mathbf{z}}_2 \mathbf{v}_{m0}) = \hat{\mathbf{z}}_2 V\mathbf{g}_{m0} + \hat{\mathbf{z}}_2 \mathbf{s}_{m0},$$

which, since the left-hand side is known (and is typically zero), forms the other linear equation to determine $\mathbf{g}_{m0}$. Once $\mathbf{g}_{m0}$ is found, (6.7) may be solved to determine $\mathbf{v}_{m0}$ up to an arbitrary multiple of the homogeneous solution $\mathbf{e}_1$. The precise multiple is determined by requiring that $\hat{\mathbf{z}}_1\mathbf{v}_{m0}$ be in accord with the definitions of x_1 and x_2 (usually requiring it to be zero).

Equation (6.6) with $n=1$, and its solution, are the same as those discussed above except that $\mathbf{v}_{m+1,0}$ appears as an additional known term of the right-hand side, and except that the roles of $\hat{\mathbf{z}}_1$ and $\hat{\mathbf{z}}_2$ are interchanged.

Equation (6.6) with $n \geqq 2$ is of the form

$$(Q+n\lambda I)\mathbf{v}_{mn} = V\mathbf{g}_{mn} + \mathbf{s}_{mn} + (m+1)\lambda\mathbf{v}_{m+1,n-1}, \tag{6.8}$$

where $\mathbf{s}_{mn}$ and $\mathbf{v}_{m+1,n-1}$ are known forcing terms. The difference here is that $Q+n\lambda I$ is, in general, not singular. In this case the two linear equations to determine $\mathbf{g}_{mn}$ come from requiring that $\mathbf{v}_{mn}$ have specific components (typically zero) of $\mathbf{e}_1$ and $\mathbf{e}_2$. Multiplying on the left by the two corresponding eigenvectors, we obtain

$$n\lambda(\hat{\mathbf{z}}_1\mathbf{v}_{mn}) = \hat{\mathbf{z}}_1 V\mathbf{g}_{mn} + \hat{\mathbf{z}}_1(\mathbf{s}_{mn} + (m+1)\lambda\mathbf{v}_{m+1,n-1}),$$

$$(n-1)\lambda(\hat{\mathbf{z}}_2\mathbf{v}_{mn}) = \hat{\mathbf{z}}_2 V\mathbf{g}_{mn} + \hat{\mathbf{z}}_2(\mathbf{s}_{mn} + (m+1)\lambda\mathbf{v}_{m+1,n-1}).$$

Upon determining $\mathbf{g}_{mn}$ we can then (almost always) solve (6.8) for $\mathbf{v}_{mn}$. This whole formal procedure for calculating the invariant manifold, given by $\mathbf{v}$, and the evolution on it, governed by $\mathbf{g}$, can be carried out to as high an order as is desired or practical.

However, if by some mischance the matrix $Q+n\lambda I$ appearing in (6.8) is singular (for $n \geqq 2$), then the asymptotic calculation can proceed no further. This is no cause for alarm, as it is equivalent to the divisions by $(1+n\lambda)$ seen and discussed in §§ 2–4. Indeed, such an occurrence (or near occurrence) warns us not to proceed, as the details at this and higher orders are then irrelevant to the invariant manifold analysis.

In summary, the procedure calculates terms in the low-dimensional asymptotic description (6.2) of the dynamical system (6.1). Typical limitations of such a description are illustrated in the discussions of the simple problem (2.1) in §§ 2–4. A remaining problem is to find the initial condition $\mathbf{x}(0)$ for the low-dimensional asymptotic description (6.2) that best matches an initial condition $\mathbf{p}(0)$ of the full system (6.1); this is discussed in Roberts [20].

7. Relation to numerical models. Many numerical solutions of a set of differential equations involve finding the behaviour of some limited number of an infinite number of complete modes. A well-known example is the set of Lorenz equations (see § 2.3 of Guckenheimer and Holmes [8]) that form an elementary numerical model of convection. The usual procedure is to decide upon a complete set of modes (typically Fourier modes), select a finite number of these with which the solution is expressed, and then substitute this into the governing equations. Identities (typically trigonometric) are used to simplify products of modes (the generation of unrepresented harmonics being ignored), and the coefficients of each mode then give equations governing the amplitude of that particular mode in the solution.

This whole conventional process can now be viewed as approximating an invariant manifold and the evolution on it. First, the subspace spanned by the retained modes (say there are N modes) is a tangent space to a corresponding N-dimensional invariant manifold, and is thus an approximation of it. Second, the evolution on this "flat" approximation to the invariant manifold is given by the equations obtained in the numerical approximation. All this leads to the possibility that such modal numerical models may be improved through this viewpoint. Instead of looking for a solution in

some "flat" N-dimensional vector space, the procedure introduced in earlier sections may be used to calculate the curving nature of the corresponding N-dimensional invariant manifold. Then the equations describing the evolution on this invariant manifold would form an improved numerical model.

Such modal numerical models obtained via this formal invariant manifold procedure would be improved in the sense that they take into account some of the dynamics of the otherwise ignored modes without explicitly including these modes in the final equations. As an example, the Lorenz equations could be modified through this approach to provide a more accurate model of convection, not by increasing the number of retained modes, but by making the three ordinary differential equations more appropriate at larger amplitudes. Thus this invariant manifold procedure is likely to be a very powerful and useful hybrid between numerical models and asymptotic expansions. How well it performs in application to the Kuramoto-Sivashinsky equations is discussed in Roberts [19].

8. Generalised applications of invariant manifolds. In almost all of the literature, the concept of an invariant manifold is applied only to dynamical systems that are dissipative, and consequently the manifold is typically stable and typically contains the large-time behaviour of the system under consideration. However, many invariant manifolds are not strictly stable but are nevertheless of immense practical importance. In this section we point out some examples that we have come across.

The simplest example of an invariant manifold that is not strictly stable occurs in any system with an (integral) invariant. Often it will be a result of conservation of mass, energy, or momentum. Such an invariant of the flow of the system restricts it to some manifold embedded in the solution space. Typically such invariants do not reduce the dimensionality of the problem significantly, but by only a few degrees of freedom. Thus there is little to be achieved by explicitly restricting the system to the corresponding invariant manifold.

An important invariant manifold in fluid mechanics is invoked by the irrotational assumption. One of the basic theorems relevant to Newtonian fluids is that if the vorticity of the fluid flow $\omega = \nabla \times \mathbf{u}$ is initially zero, then it is zero for all time. This justifies restricting attention to that class of flows that lie on the "flat" invariant manifold $\omega = 0$, that is, irrotational flow. One interesting aspect is that this approximation is usually invoked only for inviscid fluids where there is no dissipative mechanism. Thus the irrotational invariant manifold is usually not strictly stable.

An interesting class of invariant manifolds are associated with solitons. We give two examples. Burgers' equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} \tag{8.1}$$

has the $2N$-dimensional invariant manifold

$$u = -2\nu \sum_{n=1}^{N} [x - z_n(t)]^{-1}, \tag{8.2}$$

which is written in terms of N complex parameters z_n that evolve in time according to

$$\dot{z}_n = -2\nu \sum_{m \neq n} (z_n - z_m)^{-1} \tag{8.3}$$

(see Bullough and Caudrey [4]). The interesting point here is that the expressions for the invariant manifold of (8.1) and the evolution on it are exact. Furthermore, Burgers' equation is dissipative and so an invariant manifold of the above form is likely to be

stable. An invariant manifold with a similar form can be found for the Korteweg-de Vries equation that models the behaviour of long waves in many physical systems. The governing differential equation is

$$\frac{\partial u}{\partial t}+3u\frac{\partial u}{\partial x}+\frac{1}{2}\frac{\partial^3 u}{\partial x^3}=0. \tag{8.4}$$

This has the invariant manifold

$$u=2\sum_{n=1}^{N}[x-z_n(t)]^{-2} \tag{8.5}$$

(see Bullough and Caudrey [4]), on which the complex parameters evolve according to

$$\dot{z}_n=6\sum_{m\neq n}(z_n-z_m)^{-2}, \tag{8.6}$$

provided that the initial conditions for the complex parameters satisfy

$$\sum_{m\neq n}(z_m-z_n)^{-3}=0 \quad \forall n. \tag{8.7}$$

(The dimensionality of the invariant manifold is a little obscure because of the constraint that (8.7) imposes on the complex parameters z_n.) Interestingly, this invariant manifold can be used to describe the interaction of solitary waves; the conservation of the individuality of solitary waves is thus seen as a consequence of the conservation of the double poles in (8.5) as they evolve in the complex plane according to (8.6).

A similar sort of invariant manifold occurs in the two-dimensional flow of an inviscid, irrotational, and incompressible fluid. There a very important class of flows can be built up with point vortices. A $2N$-dimensional invariant manifold is described by the (scalar) vorticity field

$$\omega=\sum_{n=1}^{N}\kappa_n\delta(x-x_n)\delta(y-y_n), \tag{8.8}$$

where the locations of the vortices evolve according to

$$\dot{z}_n^*=\frac{1}{2\pi i}\sum_{\substack{m=1\\ m\neq n}}^{N}\frac{\kappa_m}{z_n-z_m}, \tag{8.9}$$

in which $z_n=x_n+iy_n$. One aspect of this invariant manifold is that it has real physical singularities. Despite this and the fact that there is no dissipation to make this an attracting manifold, it is of immense practical utility.

Other useful invariant manifolds involve neglecting oscillatory components in the full dynamical system. Thus the invariant manifold only acts as some sort of centre for the flow of the system and may best be viewed as a subcentre manifold (see Sijbrand [23, § 7] for definition and properties). Some examples are the incompressible approximation in fluid mechanics, which neglects sound waves; the quasigeostrophic approximation in atmospheric dynamics, which neglects the relatively fast gravity waves in the atmosphere (see Lorenz [10], Vautard and Legras [26], Roberts [20, § 4]); the derivation of continuum mechanics from the kinetic theory of gases, which neglects the oscillatory components in the solutions of Liouville's equations (see Muncaster [11]); and the derivation of shallow water wave equations, which neglects the oscillation of short water waves.

I imagine that all of these last four examples could be rederived and extended using the formal procedure outlined in this paper. Such derivations would provide valuable new insight into the physical nature of the approximations.

Acknowledgment. I thank Trinity College and the Department of Applied Mathematics and Theoretical Physics at the University of Cambridge for their hospitality during the preparation of this work.

REFERENCES

[1] A. ARNEODO, P. H. COULLET, AND E. A. SPIEGEL, *The dynamics of triple convection*, Geophys. Astrophys. Fluid Dynamics, 31 (1985), pp. 1-38.

[2] A. ARNEODO AND O. THOUL, *Direct numerical simulation of a triple convection problem versus the normal form prediction*, Phys. Lett. A, 109 (1985), pp. 367-373.

[3] A. J. BERNHOFF, *Transitions from order in convection*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1986.

[4] P. K. BULLOUGH AND P. J. CAUDREY, *The soliton and its history*, in Solitons, P. K. Bullough and P. J. Caudrey, eds., Topics in Current Physics 17, Springer-Verlag, Berlin, New York, 1980, pp. 1-64.

[5] J. CARR, *Applications of Centre Manifold Theory*, Appl. Math. Sci. 35, Springer-Verlag, Berlin, New York, 1981.

[6] S. C. CHIKWENDU AND G. U. OJIAKOR, *Slow-zone model for longitudinal dispersion in two-dimensional shear flows*, J. Fluid Mech., 152 (1985), pp. 15-38.

[7] P. H. COULLET AND E. A. SPIEGEL, *Amplitude equations for systems with competing instabilities*, SIAM J. Appl. Math., 43 (1983), pp. 776-821.

[8] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, Berlin, New York, 1983.

[9] G. N. MERCER AND A. J. ROBERTS, *The application of centre manifold theory to contaminant dispersion in channels and pipes with varying properties*, submitted, 1989.

[10] E. D. LORENZ, *On the existence of a slow manifold*, J. Atmospheric Sci., 43 (1986), pp. 1547-1557.

[11] R. G. MUNCASTER, *Invariant manifolds in mechanics* I: *the general construction of coarse theories from fine theories*, Arch. Rational Mech. Anal., 84 (1983), pp. 353-357.

[12] R. W. PARSON AND P. K. POLLETT, *Quasi-stationary distributions for auto-catalytic reactions*, J. Statist. Phys., 46 (1987), pp. 249-254.

[13] P. K. POLLETT AND A. J. ROBERTS, *A description of the long-term behaviour of absorbing continuous-time Markov chains using a centre manifold*, Adv. in Appl. Math., to appear.

[14] A. J. ROBERTS, *An introduction to the technique of reconstitution*, SIAM J. Math. Anal., 16 (1985), pp. 1241-1257.

[15] ———, *Simple examples of the derivation of amplitude equations possessing bifurcations*, J. Austral. Math. Soc. Ser. B, 27 (1985), pp. 48-65.

[16] ———, *The application of centre manifold theory to the evolution of systems which vary slowly in space*, J. Austral. Math. Soc. Ser. B, 29 (1988), pp. 480-500.

[17] ———, *A formal centre manifold description of the evolution of slowly-varying waves*, unpublished manuscript.

[18] ———, *The description of interacting nonlinear waves through a formal centre manifold procedure*, unpublished manuscript.

[19] ———, *The construction of invariant manifolds for the Kuramoto-Sivashinsky equation*, unpublished manuscript.

[20] ———, *Appropriate initial conditions for asymptotic descriptions of the long-term evolution of dynamical systems*, J. Austral. Math. Soc. Ser. B, 31 (1989).

[21] ———, *Effective multi-mode models of shear dispersion in channels*, in preparation.

[22] M. SHUB, *Global Stability of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1987.

[23] J. SIJBRAND, *Properties of centre manifolds*, Trans. Amer. Math. Soc., 289 (1985), pp. 431-469.

[24] R. SMITH, *Diffusion in shear flow made easy: the Taylor limit*, J. Fluid Mech., 175 (1987), pp. 201-214.

[25] G. I. TAYLOR, *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. London Ser. A, 219 (1953), pp. 186-203.

[26] R. VAUTARD AND B. LEGRAS, *Invariant manifolds, quasi-geostrophy and initialisation*, J. Atmospheric Sci., 43 (1986), pp. 565-584.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1459-1471, November 1989

011

STABILITY OF PERIODIC SOLUTIONS IN THERMOSTAT CONTROL*

GUSTAF GRIPENBERG†

Abstract. A stability criterion is given for periodic solutions of an equation describing thermostat control. The proof is based on a perturbation result for discrete functional equations.

Key words. thermostat, periodic solutions, local stability

AMS(MOS) subject classifications. 34K35, 45D05, 45M05

1. Introduction and statement of results. The purpose of this paper is to establish a criterion for the local stability of periodic solutions of the equation

$$y(t)=\int_{-\infty}^{t} a(t-s)u(s)\,ds, \qquad t\in\mathbf{R}, \tag{1}$$

where $u:\mathbf{R}\to\{0,1\}$ is determined by the requirement that if $u(t)=1$ and y reaches an upper level θ_{high} at time t, then $u(t+)=0$, and conversely, if $u(t)=0$ and y reaches a lower level θ_{low} at time t, then $u(t+)=1$.

This equation can be used to describe systems controlled by a thermostat. For example, consider a situation where a heater is turned ON if the temperature at some fixed point drops to a level θ_{low} and is turned OFF if the temperature rises to a level θ_{high}, where $\theta_{\text{low}}<\theta_{\text{high}}$. It is convenient to normalize the equation in such a way that without heating the temperature is zero and with continuous heating it is 1. The function a is determined by the requirement that if the heating is turned on at time zero for the first time, then the temperature should be $\int_0^t a(s)\,ds$ at time $t>0$. Without going into further details about how the heating process affects the temperature (e.g., through diffusion), we see that (1) can be used as a reasonable model if no external influences are involved.

In [2] and [3], where temperature regulation by thermostats is also considered, the main emphasis is on questions regarding existence of solutions and on a precise description of the heating process involving diffusion. In [5] this work is continued and some sufficient conditions for the existence of periodic solutions are given. In [1] thermostat control in a diffusion model is considered as well, and the existence, and in some cases, uniqueness, of periodic solutions is established.

In [6]–[8] a class of switching systems is considered where the state equation is of the form $y'(t)=f_j(y(t))$ and where the index j, and hence the state equation, is changed when the solution reaches a "switching surface." In [8] sufficient conditions for periodic solutions are given for this kind of system. Note that a linear system $x'(t)=Ax(t)+u(t)z$, $y(t)=Cx(t)$, can be written in the form of (1), but the analysis of (1) is not dependent on any underlying structure of this form.

In [4] the existence of periodic solutions of (1) is established, but there the dependence of u on y is in general a slightly weaker form of thermostat control (although the difference appears only in the case where y has a local maximum equal to θ_{high} or a local minimum equal to θ_{low}). Here we will use the following notion of thermostat control.

* Received by the editors January 25, 1988; accepted for publication (in revised form) November 22, 1988.
† Department of Mathematics, University of Helsinki, Regeringsgatan 15, Helsingfors, Finland.

DEFINITION. If $I \subset \mathbf{R}$ is an interval and $y: I \to \mathbf{R}$ is a continuous function, then a function $u: I \to \{0, 1\}$ is (strictly) *thermostat controlled* by y with respect to the higher limit θ_{high} and the lower limit θ_{low} on the interval I provided that u is left-continuous with right-hand limits on I, and the following conditions hold for all $t \in I$:

(i) $u(t) = 1$ if $y(t) < \theta_{\text{low}}$;

(ii) $u(t) = 0$ if $y(t) > \theta_{\text{high}}$;

(iii) If $y(t) = \theta_{\text{high}}$ and $u(t) = 1$, then $u(t+) = 0$, and if $u(t) - u(t+) = 1$, then $y(t) = \theta_{\text{high}}$;

(iv) If $y(t) = \theta_{\text{low}}$ and $u(t) = 0$, then $u(t+) = 1$, and if $u(t) - u(t+) = -1$, then $y(t) = \theta_{\text{low}}$.

Note that by this definition, the function u is constant on those intervals where $y(t) \in (\theta_{\text{low}}, \theta_{\text{high}})$.

In order to study the stability of periodic solutions we allow small perturbations and therefore we consider the equation

$$y(t) = \int_{-\infty}^{t} a(t-s)u(s)\,ds + e(t), \qquad t \geqq 0, \tag{2}$$

where u is given on $(-\infty, 0]$ and u is strictly thermostat controlled by y on $(0, \infty)$.

Since the function u can take only two values, it is completely determined by the switching times at which it jumps from zero to 1 or from 1 to zero. We will use the notation that u is 1 on the intervals $(v_n(u), w_n(u)]$ and zero on the intervals $(w_n(u), v_{n+1}(u)]$. No generality is lost if we assume that $v_0(u) = 0$. It turns out, however, that it is advantageous to use the interswitching intervals $T_n(u) = w_n(u) - v_n(u)$ and $S_n(u) = v_{n+1}(u) - w_n(u)$ instead of the switching times. Then local stability means that if e is small, $T_n(u)$ is close to some T and $S_n(u)$ is close to some S for all negative indices n; then $T_n(u)$ and $S_n(u)$ will remain close to T and S for all future (provided the system is thermostat controlled). A precise statement of what is meant by local stability is given in (• •) below. Next, we give a formal definition of $T_n(u)$ and $S_n(u)$.

DEFINITION. If $u: \mathbf{R} \to \{0, 1\}$ is left-continuous with right-hand limits, $u(0) = 0$ and $u(0+) = 1$, then the numbers $\{T_n(u)\}$ and $\{S_n(u)\}$ are defined by

$$u(t) = \begin{cases} 1, & t \in \left(-\sum_{j=n}^{-1} (T_j(u) + S_j(u)), -\sum_{j=n}^{-1} (T_j(u) + S_j(u)) + T_n \right], & n < 0, \\ 1, & t \in \left(\sum_{j=0}^{n-1} (T_j(u) + S_j(u)), \sum_{j=0}^{n-1} (T_j(u) + S_j(u)) + T_n \right], & n \geqq 0, \\ 0, & t \in \left(-\sum_{j=n}^{-1} (T_j(u) + S_j(u)) + T_{n+1}, -\sum_{j=n+1}^{-1} (T_j(u) + S_j(u)) \right], & n < 0, \\ 0, & t \in \left(\sum_{j=0}^{n-1} (T_j(u) + S_j(u)) + T_n, \sum_{j=0}^{n} (T_j(u) + S_j(u)) \right], & n \geqq 0. \end{cases}$$

It is easy to see that if u is strictly thermostat controlled by y on $(0, \infty)$, (2) holds, $\sup_{t \geqq 0} |e(t)| < \min\{\theta_{\text{low}}, 1 - \theta_{\text{high}}\}$, and if $\int_0^\infty a(t)\,dt = 1$, then $T_n(u)$ and $S_n(u)$ are defined for all $n \geqq 0$.

Since the history of the function u is determined by the values of $T_n(u)$ and $S_n(u)$ for $n < 0$, it turns out that it is possible to show that the problem of stability is reduced to a study of the stability of a discrete functional equation of the form

$$\sum_{j=-\infty}^{n} \alpha_{n-j}\xi_j = \Phi_n(\{\xi_{n-j}\}_{j=0}^{\infty}) + \varphi_n, \qquad n \geqq 0, \tag{3}$$

$$\xi_i = \psi_i, \qquad i < 0.$$

It is well known that the z-transform is useful when we study linear difference equations. The criterion for stability of the linear part of (3) is $\det[\sum_{n=0}^{\infty} z^n \alpha_n] \neq 0$ when $|z| \leqq 1$. We will see that $f(z)/(1-z)$, where f is the function appearing in (•), is exactly of this form, and that we get local stability for the nonlinear equation. (The factor $1-z$ is due to the translation invariance of the problem.)

Next we give a precise formulation of our result on the stability of periodic solutions.

THEOREM. *Assume that*

(i) $0 < \theta_{\text{low}} < \theta_{\text{high}} < 1$;

(ii) $a \in L^1(\mathbf{R}^+; \mathbf{R}) \cap AC_{\text{loc}}((0, \infty); \mathbf{R})$, $\int_0^\infty a(t)\,dt = 1$, *and* $\int_0^\infty t^2 |a'(t)|\,dt < \infty$;

(iii) $e \in C(\mathbf{R}^+; \mathbf{R})$;

(iv) $u : \mathbf{R} \to \{0, 1\}$ *is left-continuous with right-hand limits,* $u(0) = 0$, $u(0+) = 1$, *and* u *is strictly thermostat controlled by the solution* y *of* (2) *on* $(0, \infty)$;

(v) *There are positive numbers* T *and* S *such that there exists a periodic solution* (y_*, u_*) *of* (1) *with* $u_* = 1$ *on intervals of length* T, $u_* = 0$ *on intervals of length* S, $u_*(0) = 0$, $u_*(0+) = 1$, *and* u_* *is strictly thermostat controlled by* y_* *on* $\mathbf{R}$;

(vi) $y_*'(0-) < 0$ *and* $y_*(T-) > 0$.

Then the following two conditions (•) *and* (••) *are equivalent:*

(•) *The function*

$$f(z) \overset{\text{def}}{=} \sum_{j=1}^{\infty} (a(j(T+S)) - a(j(T+S)-T)) \sum_{j=1}^{\infty} (a(j(T+S)-S) - a(j(T+S)))$$

$$+ \sum_{j=1}^{\infty} (2a(j(T+S)) - a(j(T+S)-T)$$

$$- a(j(T+S)-S)) \sum_{j=1}^{\infty} a(j(T+S))z^j - \left(\sum_{j=1}^{\infty} a(j(T+S))z^j\right)^2$$

$$+ \frac{1}{z} \sum_{j=1}^{\infty} a(j(T+S)-S)z^j \sum_{j=1}^{\infty} a(j(T+S)-T)z^j$$

has a simple zero at $z = 1$ *and no other zeros in* $\{z \in \mathbf{C} \mid |z| \leqq 1\}$;

(••) *There exist numbers* $\delta > 0$ *and* $C < \infty$ *such that the inequalities*

$$|T_i(u) - T| \leqq \delta, \quad |S_i(u) - S| \leqq \delta, \quad i < 0,$$

$$|e(t)| \leqq \delta, \qquad t \geqq 0$$

imply that for every $n \geqq 0$

$$|T_n(u) - T| + |S_n(u) - S| \leqq C\left(\sup_{i<0} \{|T_i(u) - T|, |S_i(u) - S|\} + \sup_{t \geqq 0} |e(t)|\right), \tag{4}$$

and in the case where $\lim_{t\to\infty} e(t) = 0$ *it follows that*

$$T_n(u) \to T \quad \text{and} \quad S_n(u) \to S \quad \text{as } n \to \infty.$$

It is easy to see that $y_*'(0-) = \sum_{j=1}^{\infty} (a(j(T+S)) - a(j(T+S)-T))$ and $y_*'(T-) = \sum_{j=1}^{\infty} (a(j(T+S)-S) - a(j(T+S)))$. The assumption that u_* is strictly thermostat controlled by y_* implies that $y_*'(0-) \leqq 0$ and $y_*'(T-) \geqq 0$. It is not really necessary to assume that a is locally absolutely continuous, but this assumption makes the formulation of the appropriate hypotheses on the kernel a much simpler. Note also that the condition (ii) is satisfied for the kernels given in the examples in [3] and [5].

In [4, Thm. 5], it is shown that if $m \geqq 1$ is an integer and $a(t) = (1/m!)t^m e^{-t}$, then there exist positive numbers α so that there are $\lfloor (m-1)/4 \rfloor + 1$ different periodic solutions (y_*, u_*) of (1) and u_* is strictly thermostat controlled by y_* on $\mathbf{R}$ with respect to $\frac{1}{2}+\alpha$ and $\frac{1}{2}-\alpha$. (Note that there is an error in [4, Thm. 5] at this point since $\lfloor (m-1)/4 \rfloor + 1$ has been replaced there by $\lfloor m/4 \rfloor + 1$.) If the theorem above is applied in the case where $m = 5$, so that there are two periodic solutions, then numerical calculations indicate that the stability condition (•), and hence also (• •), is satisfied for the solution with the longer period only.

For completeness we state and prove the result concerning (3) that we need. Let us denote the set of integers by $\mathbf{Z}$, the set of positive integers by $\mathbf{Z}_+$, the set of negative integers by $\mathbf{Z}_-$, and the set of natural numbers, i.e., the nonnegative integers, by $\mathbf{N}$. By $|\cdot|$ we denote some norm in $\mathbf{R}^m$, and also the corresponding matrix norm.

If $\zeta : \mathbf{Z}_+ \to \mathbf{R}^m$ and $x \in \mathbf{R}^m$ then we let $[x, \zeta]$ denote the sequence given by $[x, \zeta]_j = \zeta_j$ if $j > 0$ and $[x, \zeta]_0 = x$.

PROPOSITION. *Assume that*:

(a) $\alpha \in l^1(\mathbf{N}; \mathbf{R}^{m\times m})$ *and* $\det[\alpha_0] \neq 0$;

(b) $\varphi \in l^\infty(\mathbf{N}; \mathbf{R}^m)$;

(c) *For each* $j \geqq 0$ *the mapping* $\zeta \in l^\infty(\mathbf{N}; \mathbf{R}^m) \mapsto \Phi_j(\zeta) \in \mathbf{R}^m$ *is continuous and* $\|\Phi(\zeta)\|_{l^\infty(\mathbf{N})} = o(\sup_{j\geqq 0} |\eta_j \zeta_j|)$ *as* $\|\zeta\|_{l^\infty(\mathbf{N})} \to 0$, *where* η *is some element in* $l_0^\infty(\mathbf{N}; \mathbf{R})$;

(d) $\psi \in l^\infty(\mathbf{Z}_-; \mathbf{R}^m)$;

(e) *There exist a constant* $\gamma > 0$ *and a continuous function* $\Gamma : \mathbf{R}^+ \to \mathbf{R}^+$ *with* $\Gamma(0) = 0$ *such that if* $\phi \in l^\infty(\mathbf{Z}_+; \mathbf{R}^m)$ *and* $f \in \mathbf{R}^m$ *with* $\|\phi\|_{l^\infty(\mathbf{Z}_+)} \leqq \gamma$ *and* $|f| \leqq \gamma$, *then there exists for each* $n \in \mathbf{N}$ *a unique solution of* x *of*

$$\alpha_0 x = \Phi_n([x, \phi]) + f, \tag{5}$$

and $|x| \leqq \Gamma(|f| + \|\phi\|_{l^\infty(\mathbf{Z}_+)})$.

Then the following two conditions (*) *and* (**) *are equivalent*:

(*) $\det[\sum_{n=0}^\infty z^n \alpha_n] \neq 0$ *for* $|z| \leqq 1$;

(**) *There exist numbers* $\delta > 0$ *and* $C < \infty$ *such that*

$$|\psi_i| \leqq \delta, \qquad i < 0,$$

$$|\varphi_j| \leqq \delta, \qquad j \geqq 0$$

imply that there is a unique solution ξ *of* (3), *for every* $n \geqq 0$,

$$|\xi_n| \leqq C\left(\sup_{i<0} |\psi_i| + \sup_{0\leqq k\leqq n} |\varphi_k|\right), \tag{6}$$

and, in the case where $\lim_{n\to\infty} \varphi_n = 0$, *it follows that*

$$\xi_n \to 0 \quad \text{as } n \to \infty.$$

We see from the proof of this proposition that the continuity assumption in (c) is needed only when we establish the necessity of (*). Observe that no assumptions about Lipschitz continuity are made and note also that the assumption $\det[\alpha_0] \neq 0$ in (a) actually follows from (e) (without this assumption there is clearly no hope of finding a unique solution).

If we are interested only in the stability of the solutions and do not care about the asymptotic stability, then we can take $\eta \equiv 1$ in (c).

2. Proof of the theorem. First we introduce some notation. We write $T_n = T_n(u)$, $S_n = S_n(u)$, and

$$\Delta S_n = S_n - S,$$

$$\Delta T_n = T_n - T,$$

$$v_n = \begin{cases} \sum_{j=0}^{n-1} (T_j + S_j), & n > 0, \\ -\sum_{j=n}^{-1} (T_j + S_j), & n \leqq 0, \end{cases}$$

$$w_n = v_n + T_n,$$

$$V_n = n(T+S),$$

$$W_n = V_n + T.$$

This definition says that u takes the value 1 on the intervals $(v_n, w_n]$ and zero on the intervals $(w_n, v_{n+1}]$. Similarly, $u_* = 1$ on $(V_n, W_n]$ and $u_* = 0$ on the intervals $(W_n, V_{n+1}]$. Below we will need the following relations for $n > j$:

$$\begin{aligned} &v_n + T - v_j - (V_n + T - V_j) = \sum_{i=j}^{n-1} (\Delta T_i + \Delta S_i), \\ &v_n + T - w_j - (V_n + T - W_j) = \sum_{i=j+1}^{n-1} (\Delta T_i + \Delta S_i) + \Delta S_j, \\ &w_n + S - v_j - (W_n + S - V_j) = \Delta T_n + \sum_{i=j}^{n-1} (\Delta T_i + \Delta S_i), \\ &w_n + S - w_j - (W_n + S - W_j) = \Delta T_n + \sum_{i=j+1}^{n-1} (\Delta T_i + \Delta S_i) + \Delta S_j. \end{aligned} \tag{7}$$

Since we want to obtain an equation that gives T_n as a function of T_j and $S_j, j < n$, it is advantageous to have a function that is otherwise equal to u but does not jump from 1 to zero at $v_n + T_n$. Therefore we define

$$u_\tau(t) = \begin{cases} u(t), & t \leqq \tau, \\ u(\tau+), & t > \tau. \end{cases}$$

As the number τ we will use both v_n and w_n. When we replace u by u_τ we must replace y by the function

$$y_\tau(t) = \int_{-\infty}^{t} a(t-s) u_\tau(s)\, ds. \tag{8}$$

Note that the reason for introducing these new functions is purely technical.

It is clear that T_n is determined as the smallest positive number for which $y_{v_n}(v_n + T_n) + e(v_n + T_n) = \theta_{\text{high}}$ and S_n as the smallest positive number for which $y_{w_n}(w_n + S_n) + e(w_n + S_n) = \theta_{\text{low}}$.

Now a straightforward calculation, where we use (7) and the facts that $V_n - V_j = (n-j)(T+S)$ and $V_n - W_j = (n-j)(T+S) - T$, shows that for all $n \geqq 0$ we have

$$\begin{aligned} &y_{v_n}(v_n + T) - y_*(V_n + T) \\ &\quad = \int_{-\infty}^{v_n + T} a(v_n + T - s) u_{v_n}(s)\, ds - \int_{-\infty}^{V_n + T} a(V_n + T - s) u_*(s)\, ds \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=-\infty}^{n-1} \left(\int_{v_n+T-v_j}^{v_n+T-v_j} a(s)\,ds - \int_{V_n+T-W_j}^{V_n+T-V_j} a(s)\,ds \right) \\
&= \sum_{j=-\infty}^{n-1} \left(\int_{V_n+T-V_j}^{v_n+T-v_j} a(s)\,ds - \int_{V_n+T-W_j}^{v_n+T-w_j} a(s)\,ds \right) \\
&= \sum_{j=-\infty}^{n-1} a((n-j)(T+S)+T) \sum_{i=j}^{n-1} (\Delta T_i + \Delta S_i) \qquad (9) \\
&\qquad - \sum_{j=-\infty}^{n-1} a((n-j)(T+S)) \left(\sum_{i=j+1}^{n-1} (\Delta T_i + \Delta S_i) + \Delta S_j \right) \\
&\qquad + \sum_{j=-\infty}^{n-1} \int_{V_n+T-V_j}^{v_n+T-v_j} (a(s) - a(V_n+T-V_j))\,ds \\
&\qquad - \sum_{j=-\infty}^{n-1} \int_{V_n+T-W_j}^{v_n+T-w_j} (a(s) - a(V_n+T-W_j))\,ds.
\end{aligned}
$$

We rewrite the terms on the right-hand side in (9) that are linear in ΔT_i and ΔS_i:

$$
\begin{aligned}
&\sum_{j=-\infty}^{n-1} a((n-j)(T+S)+T) \sum_{i=j}^{n-1} (\Delta T_i + \Delta S_i) \\
&\qquad - \sum_{j=-\infty}^{n-1} a((n-j)(T+S)) \left(\sum_{i=j+1}^{n-1} (\Delta T_i + \Delta S_i) + \Delta S_j \right) \\
&= \sum_{i=-\infty}^{n-1} \left(\sum_{j=n-i}^{\infty} a(j(T+S)+T) \right) (\Delta T_i + \Delta S_i) \qquad (10) \\
&\qquad - \sum_{i=-\infty}^{n-1} \left(\sum_{j=n-i+1}^{\infty} a(j(T+S)) \right) (\Delta T_i + \Delta S_i) \\
&\qquad - \sum_{i=-\infty}^{n-1} a((n-i)(T+S))\,\Delta S_i.
\end{aligned}
$$

A similar calculation shows that

$$
\begin{aligned}
&y_{w_n}(w_n+S) - y_*(W_n+S) \\
&= \int_{-\infty}^{w_n+S} a(w_n+S-s) u_{w_n}(s)\,ds - \int_{-\infty}^{W_n+S} a(W_n+S-s) u_*(s)\,ds \\
&= \sum_{j=-\infty}^{n} \left(\int_{w_n+S-w_j}^{w_n+S-v_j} a(s)\,ds - \int_{W_n+S-W_j}^{W_n+S-V_j} a(s)\,ds \right) \\
&= \sum_{j=-\infty}^{n} \left(\int_{W_n+S-V_j}^{w_n+S-v_j} a(s)\,ds - \int_{W_n+S-W_j}^{w_n+S-w_j} a(s)\,ds \right) \qquad (11) \\
&= \sum_{j=-\infty}^{n} a((n-j+1)(T+S)) \left(\sum_{i=j}^{n-1} (\Delta T_i + \Delta S_i) + \Delta T_n \right) \\
&\qquad - \sum_{j=-\infty}^{n-1} a((n-j)(T+S)+S) \left(\sum_{i=j+1}^{n-1} (\Delta T_i + \Delta S_i) + \Delta T_n + \Delta S_j \right) \\
&\qquad + \sum_{j=-\infty}^{n} \int_{W_n+S-V_j}^{w_n+S-v_j} (a(s) - a(W_n+S-V_j))\,ds \\
&\qquad - \sum_{j=-\infty}^{n} \int_{W_n+S-W_j}^{w_n+S-w_j} (a(s) - a(W_n+S-W_j))\,ds.
\end{aligned}
$$

Here the linear terms are

$$\begin{aligned}
&\sum_{j=-\infty}^{n} a((n-j+1)(T+S))\left(\sum_{i=j}^{n-1}(\Delta T_i+\Delta S_i)+\Delta T_n\right)\\
&\qquad-\sum_{j=-\infty}^{n-1} a((n-j)(T+S)+S)\left(\sum_{i=j+1}^{n-1}(\Delta T_i+\Delta S_i)+\Delta T_n+\Delta S_j\right)\\
&\quad=\sum_{i=-\infty}^{n-1}\left(\sum_{j=n-i+1}^{\infty} a(j(T+S))\right)(\Delta T_i+\Delta S_i)+\sum_{j=1}^{\infty} a(j(T+S))\,\Delta T_n \qquad (12)\\
&\qquad-\sum_{i=-\infty}^{n-1}\left(\sum_{j=n-i+1}^{\infty} a(j(T+S)+S)\right)(\Delta T_i+\Delta S_i)\\
&\qquad-\sum_{j=1}^{\infty} a(j(T+S)+S)\Delta T_n-\sum_{i=-\infty}^{n-1} a((n-i)(T+S)+S)\,\Delta S_i.
\end{aligned}$$

Since $y_*(V_n+T)=\theta_{\text{high}}=y_{v_n}(v_n+T_n)+e(v_n+T_n)=y_{v_n}(w_n)+e(w_n)$ we see that

$$\begin{aligned}
&y_{v_n}(v_n+T)-y_*(V_n+T)+y'_*(W_n-)\,\Delta T_n\\
&\quad=-\int_0^{\Delta T_n}(y'_{v_n}(w_n-s)-y'_*(W_n-))\,ds-e(w_n). \qquad (13)
\end{aligned}$$

A similar argument gives

$$\begin{aligned}
&y_{w_n}(w_n+S)-y_*(W_n+S)+y'_*(V_{n+1}-)\Delta S_n\\
&\quad=-\int_0^{\Delta S_n}(y'_{w_n}(v_{n+1}-s)-y'_*(V_{n+1}-))\,ds-e(v_{n+1}). \qquad (14)
\end{aligned}$$

From now on we assume that $e(w_n)$ and $e(v_n)$ are some given numbers, although we otherwise treat ΔT_n and ΔS_n as unknown in the equations we study.

If we combine (9) and (11) with (10) and (12), and finally also with (13) and (14), then we can rewrite the resulting two equations in vector form as

$$\sum_{j=-\infty}^{n}\alpha_{n-j}\begin{pmatrix}\Delta T_j\\ \Delta S_j\end{pmatrix}=\Phi_n\left(\left\{\begin{pmatrix}\Delta T_{n-j}\\ \Delta S_{n-j}\end{pmatrix}\right\}_{j=0}^{\infty}\right)-\begin{pmatrix}e(w_n)\\ e(v_{n+1})\end{pmatrix},\qquad n\geqq 0, \qquad (15)$$

where for each $n\geqq 0$, α_n is a 2×2 real-valued matrix, and Φ_n is a nonlinear function defined on sequences.

Now we want to apply our proposition and must verify all the assumptions.

A rather lengthy but quite straightforward calculation shows that we have

$$\det\left[\sum_{n=0}^{\infty} z^n\alpha_n\right]=\frac{f(z)}{1-z},$$

where f is the function defined in (•).

Next we consider assumption (c). Let b be a continuous nondecreasing function on $\mathbf{R}^+$ such that $b(0)=1$, $\lim_{t\to\infty} b(t)=\infty$, and $\int_0^\infty t^2 b(t)|a'(t)|\,dt<\infty$. Define the sequence $\{\eta_j\}_{j=0}^{\infty}$ by $\eta_j=1/b(j\min\{S,T\})$.

From (7) we conclude that

$$|v_n+T-v_j-(V_n+T-V_j)| \leqq \sum_{i=j}^{n-1} |\Delta T_i+\Delta S_i|$$

$$\leqq (n-j)\frac{1}{\eta_{n-j}} \sup_{i\leqq n}\{\eta_{n-i}(|\Delta T_i|+|\Delta S_i|)\}, \tag{16}$$

since the sequence η is nondecreasing and therefore $\sum_{i=j}^{n-1} 1/\eta_{n-i} \leqq (n-j)/\eta_{n-j}$.

Assume that $\sup_{i\leqq n}(|\Delta T_i|+|\Delta S_i|) \leqq \min\{S, T\}/2$. By (16) we have the following estimate:

$$\left|\sum_{j=-\infty}^{n-1}\int_{V_n+T-V_j}^{v_n+T-v_j}(a(s)-a(V_n+T-V_j))\,ds\right|$$

$$\leqq \left|\sum_{j=-\infty}^{n-1}\int_{V_n+T-V_j}^{v_n+T-v_j} a'(s)(v_n+T-v_j-s)\,ds\right|$$

$$\leqq \sup_{i\leqq n}\{\eta_{n-i}(|\Delta T_i|+|\Delta S_i|)\}$$

$$\times \sum_{j=-\infty}^{n-1}\int_{|s-(V_n+T-V_j)|\leqq\sum_{i=j}^{n-1}(|\Delta T_i|+|\Delta S_i|)} (n-j)\frac{1}{\eta_{n-j}}|a'(s)|\,ds.$$

It is clear that every term in the sum on the right-hand side in the inequality above converges to zero as $\sup_{i\leqq n}(|\Delta T_i|+|\Delta S_i|)\to 0$. To show that the sum itself goes to zero we use the dominated convergence theorem. This can be done, as we have the dominating sum

$$\sum_{j=-\infty}^{n-1}\int_{|s-(V_n+T-V_j)|\leqq(n-j)\min\{T,S\}/2}(n-j)\frac{1}{\eta_{n-j}}|a'(s)|\,ds$$

$$\leqq \frac{1}{\min\{S, T\}^2}\int_0^\infty\int_t^\infty sb(s)|a'(s)|\,ds\,dt<\infty.$$

We conclude that

$$\sum_{j=-\infty}^{n-1}\int_{V_n+T-V_j}^{v_n+T-v_j}(a(s)-a(V_n+T-V_j))\,ds$$

$$= o\left(\sup_{i\leqq n}\{\eta_{n-i}(|\Delta T_i|+|\Delta S_i|)\}\right) \quad \text{as } \sup_{i\leqq n}(|\Delta T_i|+|\Delta S_i|)\to 0.$$

In the same way we show that this conclusion holds for all the other nonlinear terms in (9) and (11) as well.

From (8) we get the following expressions for derivatives:

$$y'_{v_n}(w_n+t)=\sum_{j=-\infty}^{n-1}(a(w_n+t-v_j)-a(w_n+t-w_j))+a(w_n+t-v_n),$$

$$y'_{w_n}(v_{n+1}+t)=\sum_{j=-\infty}^{n}(a(v_{n+1}+t-v_j)-a(v_{n+1}+t-w_j)).$$

Thus we have

$$|y'_{v_n}(w_n+t)-y'_*(W_n-)|$$

$$\leqq \sum_{j=-\infty}^{n-1}|a(w_n+t-v_j)-a(W_n-V_j)|$$

$$+\sum_{j=-\infty}^{n-1}|a(w_n+t-w_j)-a(W_n-W_j)|+|a(w_n+t-v_n)-a(W_n-V_n)|.$$

Let us consider, for example, the first of these sums:

$$\sup_{|t|\leqq|\Delta T_n|}\sum_{j=-\infty}^{n-1}|a(w_n+t-v_j)-a(W_n-V_j)|$$

$$\leqq\sum_{j=-\infty}^{n-1}\int_{|s-(W_n-V_j)|\leqq 2|\Delta T_n|+\sum_{i=j}^{n-1}(|\Delta T_i|+|\Delta S_i|)}|a'(s)|\,ds.$$

It is clear that every term in this sum converges to zero as $\sup_{i\leqq n}(|\Delta T_i|+|\Delta S_i|)\to 0$, and we can again use the dominated convergence theorem and the fact that $\int_0^\infty t|a'(t)|\,dt<\infty$ to prove that the whole sum converges toward zero. All other terms can be treated in a similar manner; this gives the desired result that

$$\int_0^{\Delta T_n}(y'_{v_n}(w_n-s)-y'_*(W_n-))\,ds=o\left(\sup_{i\leqq n}\{\eta_{n-i}(|\Delta T_i|+|\Delta S_i|)\}\right),$$

$$\int_0^{\Delta S_n}(y'_{w_n}(v_{n+1}-s)-y'_*(V_{n+1}-))\,ds=o\left(\sup_{i\leqq n}\{\eta_{n-i}(|\Delta T_i|+|\Delta S_i|)\}\right)$$

as $\sup_{i\leqq n}(|\Delta T_i|+|\Delta S_i|)\to 0$.

We conclude that assumption (c) is satisfied.

Assumption (vi) implies that $\det[\alpha_0]\neq 0$ and from (13), (14), and the argument given above we can see that the existence assumption (e) is satisfied as well. (Here we use the fact that $u_*(t)<\theta_{\text{high}}$ on (V_n, W_n), and $u_*(t)>\theta_{\text{low}}$ on (W_n, V_{n+1}).)

Now an application of the proposition completes the proof. □

3. Proof of the proposition. We use the following notation: If $n\in\mathbf{Z}$ and if $\{\zeta_j\}$ is a sequence of elements in $\mathbf{R}^m$ defined at least for all indices $j\leqq n$, then the sequence $H_n\zeta:\mathbf{N}\to\mathbf{R}^m$ is defined by $(H_n\zeta)_j=\zeta_{n-j}$, i.e., $H_n\zeta$ consists of the part of ζ "before" n.

It is well known that condition (*) implies that there exists a resolvent kernel $\rho\in l^1(\mathbf{N};\mathbf{R}^m)$ such that

$$\sum_{j=0}^{n}\alpha_{n-j}\rho_j=\sum_{j=0}^{n}\rho_{n-j}\alpha_j=\delta_{0n},\qquad n\geqq 0.$$

(Here $\delta_{ij}=I$ if $i=j$ and zero otherwise.)

Choose

$$\varepsilon\in\left(0,\frac{\gamma}{1+\|\alpha\|_{l^1(\mathbf{Z}_+)}}\right]$$

to be so small that

(17) If $\|\zeta\|_{l^\infty(\mathbf{N})}\leqq\max\{\Gamma(2\varepsilon+\varepsilon\|\alpha\|_{l^1(\mathbf{Z}_+)}),\varepsilon\}$ then $\|\Phi(\zeta)\|_{l^\infty(\mathbf{N})}\leqq\dfrac{\|\zeta\|_{l^\infty(\mathbf{N})}}{2\|\rho\|_{l^1(\mathbf{N})}}$.

Take

(18) $$C=2(1+\|\alpha\|_{l^1(\mathbf{Z}_+)})\|\rho\|_{l^1(\mathbf{N})},$$

(19) $$\delta=\frac{\varepsilon}{2C}.$$

Let ψ and φ satisfy

$$|\psi_i|\leqq\delta,\qquad i<0,$$

$$|\varphi_j|\leqq\delta,\qquad j\geqq 0.$$

Assume that $n \geqq 0$ is such that (3) has a unique solution ξ_j for all $j < n$ and (6) holds with n replaced by $j < n$.

Since $C \geqq 1$, this assumption holds when $n = 0$.

Note that $|\xi_j| \leqq \varepsilon$ for all $j < n$. It follows that

$$\left| \varphi_n - \sum_{j=-\infty}^{n-1} \alpha_{n-j} \xi_j \right| < \varepsilon(1 + \|\alpha\|_{l^1(\mathbf{Z}_+)}),$$

and since (3) can be rewritten as

$$\alpha_0 \xi_n = \Phi([\xi_n, \{\xi_{n-j}\}_{j=1}^{\infty}]) + \varphi_n - \sum_{j=-\infty}^{n-1} \alpha_{n-j} \xi_j,$$

we conclude from assumption (e) that there exists a unique solution ξ_n such that $|\xi_n| \leqq \Gamma(\varepsilon(2 + \|\alpha\|_{l^1(\mathbf{Z}_+)}))$. If $|\xi_n| \leqq C(\sup_{i<0} |\psi_i| + \sup_{0 \leqq k \leqq n} |\varphi_k|)$, then the induction step is completed. Otherwise $\|H_j \xi\|_{l^\infty(\mathbf{N})} \leqq |\xi_n|$ when $j \leqq n$, so we get from (17) that

$$|\Phi_j(H_j \xi)| \leqq \frac{|\xi_n|}{2\|\rho\|_{l^1(\mathbf{N})}}, \qquad 0 \leqq j \leqq n. \tag{20}$$

With the aid of the resolvent kernel ρ we can write ξ_n as

$$\xi_n = \sum_{j=0}^{n} \rho_{n-j} \left(\Phi_j(H_j \xi) + \varphi_j - \sum_{i=-\infty}^{-1} \alpha_{j-i} \psi_i \right). \tag{21}$$

It follows from (20) and (21) that

$$|\xi_n| \leqq \tfrac{1}{2} |\xi_n| + \|\rho\|_{l^1(\mathbf{N})} \left(\sup_{0 \leqq k \leqq n} |\varphi_k| + \|\alpha\|_{l^1(\mathbf{N})} \sup_{i<0} |\psi_i| \right),$$

and we conclude that (6) holds. This completes the induction argument.

Assume next that $\lim_{n \to \infty} \varphi_n = 0$ but that $\limsup_{n \to \infty} |\xi_n| > 0$. It is clear from (6) and from our choice of δ that $|\xi_n| \leqq \varepsilon$ for all n. Hence we get from (17) that

$$\limsup_{n \to \infty} |\Phi_n(H_n \xi)| \leqq \frac{\limsup_{n \to \infty} |\xi_n|}{2\|\rho\|_{l^1(\mathbf{N})}}.$$

(Only at this point do we need the assumption that $\eta_j \to 0$; otherwise we could take $\eta_j \equiv 1$.)

On the other hand, it follows from (21) that

$$\limsup_{n \to \infty} |\xi_n| \leqq \|\rho\|_{l^1(\mathbf{N})} \limsup_{n \to \infty} |\Phi_n(H_n \xi)|$$

and combining these two inequalities we get the desired contradiction.

Next, let us show that if (*) does not hold, then (**) fails too. First we consider the case where there exists a point z_0 such that $|z_0| < 1$ and z_0 is a simple zero of the function $\det[\sum_{n=0}^{\infty} z^n \alpha_n]$ and there are no other zeros of this function in the set $\{z \mid |z| \leqq |z_0|\}$ except $\overline{z_0}$. Clearly, we may assume that $z_0 \neq 0$ and from now on we also assume that z_0 is not real. Otherwise, some small changes in the proof would be necessary.

Since z_0 is a simple zero, it follows that in a neighborhood of this point we have

$$\left[\sum_{n=0}^{\infty} z^n \alpha_n \right]^{-1} = \frac{1}{2} \frac{z_0}{z_0 - z} A + B(z), \tag{22}$$

where $A \in \mathbf{C}^{m \times m} \backslash 0$ and B is a matrix-valued function that is analytic at z_0. From (22) we immediately see that $\sum_{n=0}^{\infty} z_0^n \alpha_n A = 0$. Since all the other zeros (except $\overline{z_0}$) of

$\det[\sum_{n=0}^{\infty} z^n\alpha_n]$ have absolute value larger than $|z_0|$, there exist a number $r\in(|z_0|,1]$ and a sequence ϱ such that

$$\sum_{j=0}^{\infty} r^j|\varrho_j|<\infty, \tag{23}$$

$$\rho_j=\Re[z_0^{-j}A]+\varrho_j, \qquad j\geqq 0. \tag{24}$$

(Observe that here $\Re M$ is the matrix with each element equal to the real part of the corresponding element in M.)

Let ε be a positive number such that if $\|\zeta\|_{l^\infty(\mathbf{N})}\leqq 4\varepsilon$, then

$$\|\Phi(\zeta)\|_{l^\infty(\mathbf{N})}\leqq\|\zeta\|_{l^\infty(\mathbf{N})}\frac{1}{8}\min\left\{\frac{1}{\sum_{j=0}^{\infty}r^j|\varrho|},\frac{r-|z_0|}{|A|}\right\}. \tag{25}$$

Assume that (**) holds and let δ and C be the numbers given there. Pick a positive number q such that $4\varepsilon r^{q+1}<\min\{\delta,\varepsilon/C\}$. Let $v\in\mathbf{C}^m$ be such that $|v|=1$ and $|\Re[Av]|\geqq\frac{1}{2}|A|$.

Define the set Ω by

$$\Omega=\{\{\vartheta_j\}_{j=0}^{q}|\,\vartheta_j\in\mathbf{R}^m,|\vartheta_j|\leqq 4\varepsilon r^{q-j},j=0,1,2,\cdots,q\}$$
$$\times\left\{w\in\mathbf{C}^m\middle||w|\leqq\frac{\varepsilon}{|A|}r^q\right\}.$$

Next we consider a mapping $(\vartheta,w)\in\Omega\mapsto(G(\vartheta,w),g(\vartheta,w))$ defined as follows:

$$\begin{aligned}G_n(\vartheta,w)=&\frac{3\varepsilon}{|A|}\Re[z_0^{q-n}Av]-\sum_{j=n+1}^{q}\Re[z_0^{-n+j}A]\Phi_j(H_j\xi)\\&+\sum_{j=0}^{n}\varrho_{n-j}\Phi_j(H_j\xi),\qquad n=0,1,2,\cdots,q,\quad g(\vartheta,w)=\sum_{j=0}^{q}z_0^j\Phi_j(H_j\xi),\end{aligned} \tag{26}$$

where

$$\xi_n=\begin{cases}\Re\left[z_0^{-n}A\left(\dfrac{3\varepsilon}{|A|}z_0^q v-w\right)\right], & n<0,\\ \vartheta_n, & n=0,1,2,\cdots,q.\end{cases}$$

It is clear that the mapping $(\vartheta,w)\mapsto(G(\vartheta,w),g(\vartheta,w))$ is continuous (in the topology of $\mathbf{R}^{m\times(q+1)}\times\mathbf{C}^m$), and we must show that it maps Ω into itself. First we observe that if $(\vartheta,w)\in\Omega$, then $\sup_{j\leqq q}|\xi_j|\leqq 4\varepsilon$ and $\|H_j\xi\|_{l^\infty(\mathbf{N})}\leqq 4\varepsilon r^{q-j}$, when $0\leqq j\leqq q$. Therefore it follows from our choice of ε and the fact that $|z_0|<r\leqq 1$ that

$$\left|\sum_{j=n+1}^{q}\Re[z_0^{-n+j}A]\Phi_j(H_j\xi)\right|\leqq 4\varepsilon\frac{r-|z_0|}{8|A|}|A|\sum_{j=n+1}^{q}|z_0|^{-n+j}r^{q-j}\leqq\frac{\varepsilon}{2}r^{q-n},$$

$$\left|\sum_{j=0}^{n}\varrho_{n-j}\Phi_j(H_j\xi)\right|\leqq\frac{1}{8\sum_{j=0}^{\infty}r^j|\varrho_j|}\sum_{j=0}^{n}|\varrho_{n-j}|4\varepsilon r^{q-j}\leqq\frac{\varepsilon}{2}r^{q-n},$$

$$\left|\sum_{j=0}^{q}z_0^j\Phi_j(H_j\xi)\right|\leqq\frac{4\varepsilon(r-|z_0|)}{8|A|}\sum_{j=0}^{q}|z_0|^j r^{q-j}\leqq\frac{r^q\varepsilon}{|A|}.$$

Furthermore,

$$\left|\frac{3\varepsilon}{|A|}\Re[z_0^{q-n}Av]\right|\leqq 3\varepsilon r^{q-n},$$

and so we see that the range of the mapping in question is contained in Ω.

Since Ω is convex, we can apply Schauder's fixed-point theorem and find $(\vartheta^*, w^*) \in \Omega$ such that $G(\vartheta^*, w^*) = \vartheta^*$ and $g(\vartheta^*, w^*) = w^*$. Define ξ^* by

$$\xi_n^* = \begin{cases} \Re\left[z_0^{-n} A\left(\frac{3\varepsilon}{|A|} z_0^q v - w^*\right)\right], & n < 0, \\ \vartheta_n^*, & n = 0, 1, 2, \cdots, q. \end{cases}$$

We can now rewrite the definition of G to get

$$\xi_n^* = \Re\left[z_0^{-n} A\left(\frac{3\varepsilon}{|A|} z_0^q v - w^*\right)\right] + \sum_{j=0}^{n} (\Re[z_0^{n-j} A] + \varrho_{n-j})\Phi_j(H_j \xi^*).$$

But in view of (24) this implies that ξ^* satisfies the equation

$$\sum_{j=0}^{n} \alpha_{n-j}\xi_j = \Phi_n(H_n\xi), \qquad 0 \leqq n \leqq q,$$

$$\xi_i = \Re\left[z_0^{-i} A\left(\frac{3\varepsilon}{|A|} z_0^q v - w\right)\right], \qquad i < 0.$$

Finally, note that

$$|\xi_q^*| = |G_q(\vartheta^*, w^*)| \geqq \frac{3\varepsilon}{2} - \frac{\varepsilon}{2} = \varepsilon,$$

and we get the desired contradiction since (**) implies that $|\xi_q^*| \leqq C 4\varepsilon r^{q+1} < \varepsilon$.

If our assumption on the zero z_0 is not satisfied, but $\det[\sum_{n=0}^{\infty} z^n \alpha_n]$ has some zero with absolute value less than or equal to 1, then we proceed as follows: Assume that (**) holds and let δ and C be the positive numbers given there. Next choose a sequence $\beta \in l^1(\mathbf{N}; \mathbf{R}^m)$ such that

$$\|\alpha - \beta\|_{l^1(\mathbf{N})} < \frac{1}{16C}$$

and such that $\det[\sum_{n=0}^{\infty} z^n \beta_n]$ has a simple zero in some point z_0 with $|z_0| < 1$ and no other zeros except $\overline{z_0}$, with absolute value less than or equal to $|z_0|$. It is not too difficult to see that this can always be done.

As we have seen in the above, we can find a sequence $\psi \in l^\infty(\mathbf{Z}_-; \mathbf{R}^m)$ and an integer $q > 0$ such that $\sup_{i<0} |\psi_i|$ is arbitrarily small and the solution ξ^* of the equation

$$\sum_{j=0}^{n} \beta_{n-j}\xi_j = \Phi_n(H_n\xi), \qquad 0 \leqq n \leqq q,$$

$$\xi_i = \psi_i, \qquad i < 0$$

satisfies $|\xi_j^*| \leqq 16C \sup_{i<0} |\psi_i|$, $0 \leqq j \leqq q$, and $|\xi_q^*| \geqq 4C \sup_{i<0} |\psi_i|$. But then ξ^* is also a solution of the equation

$$\sum_{j=0}^{n} \alpha_{n-j}\xi_j = \Phi_n(H_n\xi) + \varphi_n, \qquad 0 \leqq n \leqq q,$$

$$\xi_i = \psi_i, \qquad i < 0,$$

where $\varphi_n = \sum_{j=0}^{n} (\alpha_{n-j} - \beta_{n-j})\xi_j^*$. But now $|\varphi_j| \leqq 16C \sup_{i<0} |\psi_i|/(16C)$, $0 \leqq j \leqq q$, and since $|\xi_q| \geqq 4C \sup_{i<0} |\psi_i^*|$ we get a contradiction. □

Acknowledgment. I thank the referees for several helpful suggestions.

REFERENCES

[1] A. FRIEDMAN AND L.-S. JIANG, *Periodic solutions for a thermostat control problem*, Comm. Partial Differential Equations, 13 (1988), pp. 515–550.

[2] K. GLASHOFF AND J. SPREKELS, *An application of Glicksberg's theorem to set-valued integral equations arising in the theory of thermostats*, SIAM J. Math. Anal., 12 (1981), pp. 477–486.

[3] ———, *The regulation of temperature by thermostats and set-valued integral equations*, J. Integral Equations, 4 (1982), pp. 95–112.

[4] G. GRIPENBERG, *On periodic solutions of a thermostat equation*, SIAM J. Math. Anal., 18 (1987), pp. 694–702.

[5] J. PRÜSS, *Periodic solutions of the thermostat problem*, in Proc. Conference on Differential Equations in Banach Spaces, Bologna, July 1985, Lecture Notes in Math. 1223, Springer-Verlag, Berlin, New York, 1987, pp. 216–226.

[6] T. I. SEIDMAN, *Optimal control for switching systems*, in Proc. Conference on Information Science and Systems, The Johns Hopkins University, Baltimore, MD, 1987, pp. 485–489.

[7] ———, *Switching systems*, I, Mathematical Research Report 86-78, University of Maryland, Baltimore, MD, 1986.

[8] ———, *Switching systems and periodicity*, Mathematical Research Report 88-02, University of Maryland, Baltimore, MD, 1988.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1472-1483, November 1989

012

FAR-FIELD PATTERNS FOR ACOUSTIC WAVES IN AN INHOMOGENEOUS MEDIUM*

DAVID COLTON†, ANDREAS KIRSCH‡, AND LASSI PÄIVÄRINTA§

Abstract. This paper is concerned with the class of far-field patterns corresponding to the scattering of time-harmonic acoustic plane waves by an inhomogeneous medium of compact support. This class is shown to be complete in $L^2(\partial\Omega)$ (where $\partial\Omega$ is the unit sphere) for any positive value of the wave number, with the possible exception of a discrete set of wave numbers called transmission eigenvalues. The existence of a unique weak solution to the interior transmission problem (which plays a basic role in a new method for solving the inverse scattering problem) is also established for any positive value of the wave number provided the wave number is not a transmission eigenvalue.

Key words. far-field patterns, acoustic waves, inverse scattering

AMS(MOS) subject classifications. 35P25, 76Q05

1. Introduction. In this paper, as in a previous paper by Colton [1], we are concerned with the class of far-field patterns corresponding to the scattering of time-harmonic acoustic plane waves by an inhomogeneous medium of compact support. In [1] it is shown that this class of far-field patterns is complete in $L^2(\partial\Omega)$, where $\partial\Omega$ is the unit sphere provided that the wave number is sufficiently small. The purpose of this paper is to remove this restriction on the wave number and show that the class of far-field patterns is complete in $L^2(\partial\Omega)$ for any positive value of the wave number with the possible exception of a discrete set of wave numbers called *transmission eigenvalues.* (In the case of a spherically symmetric medium it is shown in [3] that these transmission eigenvalues actually exist and can be numerically determined from the far-field data.) We will also consider a recently introduced boundary value problem for the reduced wave equation called the *interior transmission problem,* which plays a basic role in a new method for solving the inverse scattering problem introduced by Colton and Monk in [3]. In [3] it is shown that there exists a unique weak solution to the interior transmission problem provided the wave number is sufficiently small. Here we will establish the existence of a unique weak solution to the interior transmission problem for any positive value of the wave number provided the wave number is not a transmission eigenvalue.

In passing, we would like to note the recent paper of Kirsch [6], which also studies the relationship between far-field patterns and the interior transmission problem for acoustic waves, as well as the work of Colton and Päivärinta [4], which considers the case of electromagnetic wave propagation in an inhomogeneous medium.

2. Wave propagation in an inhomogeneous medium and far-field patterns. Consider the scattering due to a nonabsorbing inhomogeneous medium of compact support of

* Received by the editors October 19, 1988; accepted for publication March 6, 1989.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716. The research of this author was supported in part by the Air Force Office of Scientific Research, the National Science Foundation, and the Deutsche Forschungsgemeinschaft.

‡ Institut für Angewandte Mathematik, Universität Erlangen, Erlangen, Federal Republic of Germany. The research of this author was supported in part by the Deutsche Forschungsgemeinschaft.

§ Department of Mathematics, University of Helsinki, Helsinki, Finland. The research of this author was supported in part by the Academy of Finland.

the incident plane wave

$$u^i(\mathbf{x}, t) \equiv \exp[ik\mathbf{x}\cdot\hat{\alpha} - i\omega t] \tag{2.1}$$

where $k>0$ is the wave number, ω is the frequency, and $\hat{\alpha}$, $|\hat{\alpha}|=1$, is the direction of propagation. Let $c(\mathbf{x})$, $\mathbf{x}\in R^3$, denote the local speed of sound and assume that $c(\mathbf{x})=c_0>0$ for $r=|\mathbf{x}|>a$ where c_0 is a constant. Then, if $k=\omega/c_0>0$, $n(\mathbf{x})=(c_0/c(\mathbf{x}))^2$, and we factor out the term $e^{-i\omega t}$, under appropriate assumptions (cf. [8]) the mathematical problem we are faced with is to determine the velocity potential $u(\mathbf{x})$ of the total field such that

$$\Delta_3 u + k^2 n(\mathbf{x}) u = 0 \quad \text{in } R^3, \tag{2.2}$$

$$u(\mathbf{x}) \equiv \exp[ik\mathbf{x}\cdot\hat{\alpha}] + u^s(\mathbf{x}), \tag{2.3}$$

$$\lim_{r\to\infty} r\left(\frac{\partial u^s}{\partial r} - iku^s\right) = 0 \tag{2.4}$$

where $u^s(\mathbf{x})$ denotes the scattered field, and the Sommerfeld radiation condition (2.4) is assumed to hold uniformly for $\hat{\mathbf{x}}=\mathbf{x}/|\mathbf{x}|$ on the unit sphere $\partial\Omega$. As in [1] and [3], we will make the assumption that $n(\mathbf{x})$ is positive and continuously differentiable and that

$$B=\{\mathbf{x}\in R^3: n(\mathbf{x})\neq 1\} \tag{2.5}$$

is simply connected (with C^2 boundary ∂B) and contains the origin. In particular, this implies that for $\mathbf{x}\in B$, either $c(\mathbf{x})>c_0$ or $0<c(\mathbf{x})<c_0$.

The scattering problem (2.2)-(2.4) is easily seen to be equivalent to the integral equation

$$u(\mathbf{x}) = \exp[ik\mathbf{x}\cdot\hat{\alpha}] - k^2\int_B\int \Phi(\mathbf{x},\mathbf{y})m(\mathbf{y})u(\mathbf{y})\,d\mathbf{y} \tag{2.6}$$

where $u(\mathbf{x})=u(\mathbf{x}; k, \hat{\alpha})$,

$$m(\mathbf{x}) = 1 - n(\mathbf{x}) \tag{2.7}$$

and

$$\Phi(\mathbf{x},\mathbf{y}) = \frac{\exp[ik|\mathbf{x}-\mathbf{y}|]}{4\pi|\mathbf{x}-\mathbf{y}|}. \tag{2.8}$$

From (2.3) and (2.6) it is seen that $u^s(\mathbf{x})$ has the asymptotic behavior

$$u^s(\mathbf{x}) = \frac{e^{ikr}}{r}F(\hat{\mathbf{x}}; k, \hat{\alpha}) + O\left(\frac{1}{r^2}\right) \tag{2.9}$$

where

$$F(\hat{\mathbf{x}}; k, \hat{\alpha}) = -\frac{k^2}{4\pi}\int_B\int \exp[-ik\hat{\mathbf{x}}\cdot\mathbf{y}]m(\mathbf{y})u(\mathbf{y})\,d\mathbf{y}. \tag{2.10}$$

The function $F(\hat{\mathbf{x}}; k, \hat{\alpha})$ is known as the *far-field pattern* corresponding to the incident plane wave (2.1). Noting that for $\mathbf{x}\in B$, $m(\mathbf{x})$ is either always positive or always negative, we assume without loss of generality that $m(\mathbf{x})$ is positive for $\mathbf{x}\in B$ and define the Hilbert space $L^2_m(B)$ by

$$L^2_m(B) = \left\{u: u \text{ measurable}, \int_B\int m|u|^2\,d\mathbf{x} < \infty\right\} \tag{2.11}$$

with the inner product given by

$$(f, g) = \int_B \int mf\bar{g}\, d\mathbf{x}. \tag{2.12}$$

(If $m(\mathbf{x})$ is negative for $\mathbf{x} \in B$, then $m(\mathbf{x})$ must be replaced by $-m(\mathbf{x})$ in the definition above.) We now define the compact operator $\mathbf{T}_k : L^2_m(B) \to L^2_m(B)$ by

$$\mathbf{T}_k u = \int_B \int \Phi(\mathbf{x}, \mathbf{y}) m(\mathbf{y}) u(\mathbf{y})\, d\mathbf{y} \tag{2.13}$$

and write (2.6) in the short form

$$u = \exp[ik\mathbf{x} \cdot \hat{\alpha}] - k^2 \mathbf{T}_k u. \tag{2.14}$$

The existence of a unique solution to (2.6) (or 12.14) is established in [8] and [3].

Let $\{\hat{\alpha}_n\}_1^\infty$ be a countable dense set of vectors on the unit sphere $\partial\Omega$ and for each fixed $k > 0$ define the class $\mathbf{F}$ of far-field patterns by

$$\mathbf{F} = \{F(\hat{\mathbf{x}}; k, \hat{\alpha}_n) : n = 1, 2, 3, \cdots\}. \tag{2.15}$$

Our aim is to show that, except possibly for a discrete set of values of $k > 0$, the set $\mathbf{F}$ is complete in $L^2(\partial\Omega)$. To do this we will need four lemmas. The first of these is the following *reciprocity principle.*

LEMMA 1. *Let $F(\hat{\mathbf{x}}; k, \hat{\alpha})$ be the far-field pattern corresponding to the incident plane wave* (2.1). *Then*

$$F(\hat{\mathbf{x}}; k, \hat{\alpha}) = F(-\hat{\alpha}; k, -\hat{\mathbf{x}}).$$

Proof. Using the representation theorem for the scattered field u^s [2] and the asymptotic behavior of the fundamental solution $\Phi(\mathbf{x}, \mathbf{y})$, we have

$$\begin{aligned}
F(\hat{\mathbf{x}}; k, \hat{\alpha}) &= \frac{1}{4\pi} \int_{\partial B} \left\{ u^s(\mathbf{y}, k, \hat{\alpha}) \frac{\partial}{\partial \nu_y} \exp[-ik\hat{\mathbf{x}} \cdot \mathbf{y}] \right. \\
&\qquad \left. - \exp[-ik\hat{\mathbf{x}} \cdot \mathbf{y}] \frac{\partial}{\partial \nu} u^s(\mathbf{y}; k, \hat{\alpha}) \right\} ds(\mathbf{y}) \\
&= \frac{1}{4\pi} \int_{\partial B} \left\{ u^s(\mathbf{y}; k, \hat{\alpha}) \frac{\partial}{\partial \nu} (u(\mathbf{y}; k, -\hat{\mathbf{x}}) - u^s(\mathbf{y}; k, -\hat{\mathbf{x}})) \right. \\
&\qquad \left. - (u(\mathbf{y}; k, -\hat{\mathbf{x}}) - u^s(\mathbf{y}; k, -\hat{\mathbf{x}})) \frac{\partial}{\partial \nu} u^s(\mathbf{y}; k, \hat{\alpha}) \right\} ds(\mathbf{y}) \\
&= \frac{1}{4\pi} \int_{\partial B} \left\{ (u(\mathbf{y}; k, \hat{\alpha}) - \exp[ik\mathbf{y} \cdot \hat{\alpha}]) \frac{\partial}{\partial \nu} u(\mathbf{y}; k, -\hat{\mathbf{x}}) \right. \\
&\qquad \left. - u(\mathbf{y}; k, -\hat{\mathbf{x}}) \frac{\partial}{\partial \nu} (u(\mathbf{y}; k, \hat{\alpha}) - \exp[ik\mathbf{y} \cdot \hat{\alpha}]) \right\} ds(\mathbf{y}) \\
&= \frac{-1}{4\pi} \int_{\partial B} \left\{ \exp[ik\mathbf{y} \cdot \hat{\alpha}] \frac{\partial}{\partial \nu} u(\mathbf{y}; k, -\hat{\mathbf{x}}) \right. \\
&\qquad \left. - u(\mathbf{y}; k, -\hat{\mathbf{x}}) \frac{\partial}{\partial \nu} \exp[ik\mathbf{y} \cdot \hat{\alpha}] \right\} ds(\mathbf{y}) \\
&= F(-\hat{\alpha}; k, -\hat{\mathbf{x}})
\end{aligned} \tag{2.16}$$

where ν is the unit outward normal to ∂B and we use the identities

$$\begin{aligned}&\int_{\partial B}\left\{u^s(\mathbf{y};k,\hat{\alpha})\frac{\partial}{\partial\nu}u^s(\mathbf{y};k,-\hat{\mathbf{x}})-u^s(\mathbf{y};k,-\hat{\mathbf{x}})\frac{\partial}{\partial\nu}u^s(\mathbf{y};k,\hat{\alpha})\right\}ds(\mathbf{y})=0,\\&\int_{\partial B}\left\{u(\mathbf{y};k,\hat{\alpha})\frac{\partial}{\partial\nu}u(\mathbf{y};k,-\hat{\mathbf{x}})-u(\mathbf{y};k,-\hat{\mathbf{x}})\frac{\partial}{\partial\nu}u(\mathbf{y};k,\hat{\alpha})\right\}ds(\mathbf{y})=0,\end{aligned}\tag{2.17}$$

which follow from Green's theorem.

LEMMA 2. *The orthogonal complement of* $\mathbf{F}$ *in* $L^2(\partial\Omega)$ *consists of those functions* $g\in L^2(\partial\Omega)$ *for which there exists* $w\in C^2(B)\cap C^1(\bar{B})$ *and* v *defined by*

$$v(\mathbf{x})=\int_{\partial\Omega}g(\hat{\mathbf{y}})\exp[ik\mathbf{x}\cdot\hat{\mathbf{y}}]\,ds(\hat{\mathbf{y}})\tag{2.18}$$

such that $\{v, w\}$ *is a solution to*

$$\begin{aligned}&\Delta_3 w+k^2n(\mathbf{x})w=0 && \text{in } B,\\&\Delta_3 v+k^2v=0 && \text{in } B,\\&w=v && \text{on } \partial B,\\&\frac{\partial w}{\partial\nu}=\frac{\partial v}{\partial\nu} && \text{on } \partial B.\end{aligned}\tag{2.19}$$

Remark. Functions v of the form (2.18) are called *Herglotz wave functions* with *Herglotz kernel* g (cf. [5]). The boundary value problem (2.19) is the (homogeneous) *interior transmission problem* first studied in [3].

Proof. Let $\mathbf{F}^\perp$ denote the orthogonal complement to $\mathbf{F}$. We will show that if $g\in\mathbf{F}^\perp$ then g satisfies the assumptions stated in the lemma. The fact that, if g satisfies the assumptions of the lemma, then $g\in\mathbf{F}^\perp$, follows from a simple application of Green's formula.

Suppose $g\in\mathbf{F}^\perp$, i.e.,

$$\int_{\partial\Omega}F(\hat{\mathbf{x}};k,\hat{\alpha}_n)\overline{g(\hat{\mathbf{x}})}\,ds(\hat{\mathbf{x}})=0\tag{2.20}$$

for $n=1,2,\cdots$. From Lemma 1 and the continuity of F as a function of $\hat{\alpha}$ we have that

$$\int_{\partial\Omega}F(-\hat{\alpha};k,-\hat{\mathbf{x}})\overline{g(\hat{\mathbf{x}})}\,ds(\hat{\mathbf{x}})=0\tag{2.21}$$

for all $\hat{\alpha}\in\partial\Omega$, i.e.,

$$\int_{\partial\Omega}F(\hat{\mathbf{x}};k,\hat{\alpha})\overline{g(-\hat{\alpha})}\,ds(\hat{\alpha})=0\tag{2.22}$$

for all $\hat{\mathbf{x}}\in\partial\Omega$. By superposition we note that the left-hand side of (2.22) is the far-field pattern of the scattered field w^s corresponding to the incident field

$$\begin{aligned}w^i(\mathbf{x})&=\int_{\partial\Omega}\overline{g(-\hat{\alpha})}\exp[ik\mathbf{x}\cdot\hat{\alpha}]\,ds(\hat{\alpha})\\&=\int_{\partial\Omega}g(\hat{\alpha})\exp[ik\mathbf{x}\cdot\hat{\alpha}]\,ds(\hat{\alpha}).\end{aligned}\tag{2.23}$$

But from (2.22) we have that the far-field pattern of w^s vanishes and hence w^s vanishes in $R^3\backslash B$, i.e., if $w = w^s + w^i$, then $w = w^i$ on ∂B and $\partial w/\partial \nu = \partial w^i/\partial \nu$ on ∂B. Taking the conjugate of w now implies the lemma.

To show that $\mathbf{F}$ is complete in $L^2(\partial\Omega)$ except for possibly a discrete set of k values, it now suffices by Lemma 2 and the theory of Herglotz wave functions [5] to show that the eigenvalues of (2.19) form a discrete set, i.e., except for possibly a discrete set of k values the only solution of (2.19) is $\{v, w\}$ identically zero. To this end we define the discrete set I by

$$I = \{k: k^2 \text{ is a Dirichlet eigenvalue of } -\Delta_3 \text{ in } B\}$$

and define the vector space W by

$$W = \left\{ u \in C^2(B) \cap C^1(\bar{B}): u = \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial B, \int_B \int \frac{1}{m} (|u|^2 + |\Delta_3 u|^2)\, d\mathbf{x} < \infty \right\}.$$

We make W a Hilbert space $\bar{W}$ by defining on W the inner product

$$\langle u, v\rangle = \int_B \int \frac{1}{m} (u\bar{v} + \Delta_3 u \Delta_3 \bar{v})\, d\mathbf{x}, \qquad u, v \in W \tag{2.24}$$

and completing W with respect to the norm $|u| = \sqrt{\langle u, u\rangle}$.

We first want to define an isomorphism $\mathbf{S}(k)$ on $\bar{W}$ that depends analytically on k for k in a complex neighborhood of $k_0 \in R^+$. To do this, we let $G(\mathbf{x}, \mathbf{y})$ be the Dirichlet Green's function for $\Delta_3 + k^2$ in B, $k \in R^+\backslash I$, and make the assumption that there exists a positive constant $\gamma = \gamma(k)$ such that for all $\mathbf{y}, \mathbf{z} \in \bar{B}$ we have

$$\left| \int_B \int \frac{\sqrt{m(\mathbf{y})m(\mathbf{z})}}{m(\mathbf{x})} G(\mathbf{x}, \mathbf{y}) G(\mathbf{x}, \mathbf{z})\, d\mathbf{x} \right| \leqq \gamma(\mathrm{k}) \tag{2.25}$$

where the integral in (2.25) is uniformly convergent. Roughly speaking, condition (2.25) says that, if $m(\mathbf{x})$ is continuously differentiable, then for $\mathbf{y} \in \partial B$, $m(\mathbf{x}) = O(|\mathbf{x} - \mathbf{y}|^\alpha)$, $1 \leqq \alpha < 3$, as $\mathbf{x} \to \mathbf{y}$. We can now prove the following two lemmas.

LEMMA 3. *Assume that* (2.25) *is valid. Then for all* $k_0 \in R^+$ *there exists a positive constant* $c = c(k_0)$ *such that*

$$|u|^2 \leqq c(k_0) \int_B \int \frac{1}{m} |\Delta_3 u + k_0^2 u|^2\, d\mathbf{x}$$

for $u \in \bar{W}$.

Proof. We first choose $k_0 \in R^+\backslash I$. Then for $u \in \bar{W}$ we have the representation

$$u(\mathbf{x}) = -\int_B \int (\Delta_3 + k_0^2) u(\mathbf{y}) G(\mathbf{x}, \mathbf{y})\, d\mathbf{y}, \qquad \mathbf{x} \in B \tag{2.26}$$

and hence we have

$$\begin{aligned} \int_B \int \frac{1}{m} |u|^2\, d\mathbf{x} = \int_B \int \int_B \int (\Delta_3 + k_0^3) u(\mathbf{y}) (\Delta_3 + k_0^2) \overline{u(\mathbf{z})} \\ \cdot \int_B \int \frac{1}{m(\mathbf{x})} G(\mathbf{x}, \mathbf{y}) G(\mathbf{x}, \mathbf{y}) G(\mathbf{x}, \mathbf{z})\, d\mathbf{x}\, d\mathbf{y}\, d\mathbf{z}. \end{aligned} \tag{2.27}$$

We now see that

$$\begin{aligned} \int_B \int \frac{1}{m} |u|^2\, d\mathbf{x} &\leqq \gamma(k_0) \left| \int_B \int \frac{1}{\sqrt{m}} (\Delta_3 + k_0^2) u\, d\mathbf{x} \right|^2 \\ &\leqq \gamma(k_0)\, \mathrm{vol}\, B \int_B \int \frac{1}{m} |(\Delta_3 + k_0^2) u|^2\, d\mathbf{x}. \end{aligned} \tag{2.28}$$

In addition,

$$\int_B \int \frac{1}{m} |\Delta_3 u|^2 \, d\mathbf{x} \leqq 2 \int_B \int \frac{1}{m} |\Delta_3 u + k_0^2 u|^2 \, d\mathbf{x} + 2k_0^2 \int_B \int \frac{1}{m} |u|^2 \, d\mathbf{x} \tag{2.29}$$

and (2.28), (2.29) imply the lemma is true for $k_0 \in R^+ \backslash I$.

Now assume $k_0 \in I$ and let $k_1 \in R^+ \backslash I$. From the analysis above, we see that $\Delta_3 + k_1^2 \colon \bar{W} \to L^2_{1/m}(B)$ is injective and has closed range, i.e., $\Delta_3 + k_1^2$ is a semi-Fredholm operator. But $\Delta_3 + k_0^2 = \Delta_3 + k_1^2 + (k_0^2 - k_1^2)$ is also semi-Fredholm since from (2.25) and Rellich's lemma I: $\bar{W} \hookrightarrow L^2_{1/m}(B)$ is easily seen to be compact (a bit more effort shows that this can also be established without assuming that (2.25) is uniformly convergent). Since $\Delta_3 + k_0^2$ is injective (represent u as in (2.26) with $G(\mathbf{x}, \mathbf{y})$ replaced by $\Phi(\mathbf{x}, \mathbf{y})$ and note that the norm in $L^2_{1/m}(B)$ dominates the norm in $L^2(B)$) we can now conclude that the lemma is true for $k = k_0$ since the range of a semi-Fredholm operator is closed.

The lemma is now proved.

Now fix $k_0 \in R^+$ and define for $u, v \in \bar{W}$ the inner product

$$\langle u, v \rangle_{k_0} = \int_B \int \frac{1}{m} (\Delta_3 u + k_0^2 u)(\Delta_3 \bar{v} + k_0^2 \bar{v}) \, d\mathbf{x} \tag{2.30}$$

with norm $|u|_{k_0} = \sqrt{\langle u, v \rangle_{k_0}}$. By Lemma 3 the norm $|\cdot|_{k_0}$ is equivalent to $|\cdot|$ for any $k_0 \in R^+$.

LEMMA 4. *For arbitrary complex k define the sesquilinear form* $\mathbf{B}$ *on* $\bar{W}$ *by*

$$\mathbf{B}(u, v; k) = \int_B \int \frac{1}{m} (\Delta_3 + k^2) u (\Delta_3 + k^2) \bar{v} \, d\mathbf{x}$$

where it is assumed that (2.25) *is valid. Then for every $k_0 \in R^+$ there exists $\varepsilon > 0$ such that if $|k - k_0| < \varepsilon$ then*

$$|\mathbf{B}(u, v; k) - \mathbf{B}(u, v; k_0)| \leqq C |u|_{k_0} |v|_{k_0}$$

where C is a constant satisfying $0 < C < 1$.

Proof. $\mathbf{B}$ is well defined by Lemma 3. We have

$$\begin{aligned} \mathbf{B}(u, v; k) - \mathbf{B}(u, v; k_0) &= (k^2 - k_0^2) \int_B \int \frac{1}{m} (u \Delta_3 \bar{v} + \bar{v} \Delta_3 u) \, d\mathbf{x} \\ &\quad + (k^4 - k_0^4) \int_B \int \frac{1}{m} u \bar{v} \, d\mathbf{x} \end{aligned} \tag{2.31}$$

and hence by Schwarz's inequality

$$\begin{aligned} |\mathbf{B}(u, v; k) - \mathbf{B}(u, v; k_0)| &\leqq |k^2 - k_0^2| \left(\int_B \int \frac{1}{m} |u|^2 \, d\mathbf{x} \right)^{1/2} \left(\int_B \int \frac{1}{m} |\Delta_3 v|^2 \, d\mathbf{x} \right)^{1/2} \\ &\quad + |k^2 - k_0^2| \left(\int_B \int \frac{1}{m} |v|^2 \, d\mathbf{x} \right)^{1/2} \left(\int_B \int \frac{1}{m} |\Delta_3 u|^2 \, d\mathbf{x} \right)^{1/2} \\ &\quad + |k^4 - k_0^4| \left(\int_B \int \frac{1}{m} |u|^2 \, d\mathbf{x} \right)^{1/2} \left(\int_B \int \frac{1}{m} |v|^2 \, d\mathbf{x} \right)^{1/2}. \end{aligned} \tag{2.32}$$

From Lemma 3 we now have that

$$|\mathbf{B}(u, v; k) - \mathbf{B}(u, v; k_0)| \leqq (2|k^2 - k_0^2| + |k^4 - k_0^4|) c(k_0) |u|_{k_0} |v|_{k_0}. \tag{2.33}$$

Hence if $|k - k_0|$ is sufficiently small, then $(2|k^2 - k_0^2| + |k^4 - k_0^4|) c(k_0)$ is less than 1 and the lemma follows.

Since $\mathbf{B}(u, v; k_0) = \langle u, v\rangle_{k_0}$, we see from Lemma 4 that for each $v \in \bar{W}$, $\mathbf{B}(u, v; k)$ is a bounded linear functional on $\bar{W}$. Hence, by the Riez representation theorem, there exists $v_k \in \bar{W}$ such that

$$\mathbf{B}(u, v; k) = \langle u, v_k\rangle_{k_0}. \tag{2.34}$$

Since $v \mapsto v_k$ is linear, there exists a linear operator $\mathbf{T} = \mathbf{T}(k)$ such that

$$\mathbf{B}(u, v; k) = \langle u, \mathbf{T}(k)v\rangle_{k_0} \tag{2.35}$$

for all $u, v \in \bar{W}$. Since from Lemma 4 we have that

$$|\mathbf{B}(u, v; k)| \leqq (1 + C)|u|_{k_0}, \tag{2.36}$$

we see that $\mathbf{T}(k)$ is a bounded operator. Now let $\mathbf{S}(k)$ be the adjoint of $\mathbf{T}(k)$ in W, i.e.,

$$B(u, v; k) = \langle \mathbf{S}(k)u, v\rangle_{k_0}. \tag{2.37}$$

We claim that $\mathbf{S}(k)$ is an isomorphism on $\bar{W}$ that depends analytically on k. The analyticity follows from the theory of operator-valued analytic functions [7] since from (2.37), $\mathbf{S}(k)$ is weakly analytic and hence strongly analytic. $\mathbf{S}(k)$ is injective, since if $\mathbf{S}(k)u = 0$, then $\mathbf{B}(u, v; k) = 0$ for every $v \in \bar{W}$, and hence from Lemma 4 we have $\mathbf{B}(u, u; k_0) = 0$, which implies that $u = 0$. The range of $\mathbf{S}(k)$ is dense in W, since if $\langle \mathbf{S}(k)u, v\rangle_{k_0} = 0$ for every $u \in \bar{W}$, then by (2.37), $\mathbf{B}(u, v; k) = 0$ for every $u \in \bar{W}$ and from Lemma 4 we can conclude that $v = 0$. Hence, to show that $\mathbf{S}(k)$ is an isomorphism, we now only need to show that the range of $\mathbf{S}(k)$ is closed. It suffices to show that there exists $\delta > 0$ such that for every $u \in \bar{W}$,

$$|\mathbf{S}(k)u|_{k_0} \geqq \delta |u|_{k_0} \tag{2.38}$$

and we need only consider u such that $|u|_{k_0} = 1$. For such a u we have

$$\begin{aligned} |\mathbf{S}(k)u|_{k_0} &= \sup_{|v|_{k_0} \leqq 1} |\langle v, \mathbf{S}(k)u\rangle_{k_0}| \\ &= \sup_{|v|_{k_0} \leqq 1} |\mathbf{B}(u, v; k)| \\ &\geqq |\mathbf{B}(u, u; k)|, \end{aligned} \tag{2.39}$$

$$\begin{aligned} |\mathbf{B}(u, u; k)| &\geqq \mathbf{B}(u, u; k_0) - |\mathbf{B}(u, u; k) - \mathbf{B}(u, u; k_0)| \\ &\geqq |u|_{k_0}^2 - C|u|_{k_0}^2 \\ &= 1 - C. \end{aligned} \tag{2.40}$$

Inequality (2.38) now follows from (2.39) and (2.40), recalling that $|u|_{k_0} = 1$.

We are now in a position to prove the main result of this section.

THEOREM 1. *Assume that* (2.25) *is valid. Then, except possibly for a discrete set of values of* $k > 0$, *the set* $\mathbf{F}$ *is complete in* $L^2(\partial\Omega)$.

Proof. We began by defining a projection operator $\mathbf{P}_k$ in $L_m^2(B)$ depending on $k \in R^+$. Let $f \in L_m^2(B)$ and for fixed $k_0 \in R^+$ define the linear functional l_f by

$$l_f(\phi) = \int_B \int \bar{f}(\Delta_3 + k^2)\mathbf{S}^{-1}(k)\phi \, d\mathbf{x} \tag{2.41}$$

where $\phi \in \bar{W}$ and k is such that $|k - k_0| < \varepsilon$, where ε is defined as in Lemma 4. Then l_f is bounded on $\bar{W}$ by Lemma 3 and the fact that $\mathbf{S}^{-1}(k)$ is bounded. By the Riesz representation theorem, there exists $p_f \in \bar{W}$ such that for all $\phi \in \bar{W}$

$$l_f(\phi) = \langle \phi, p_f\rangle_{k_0} \tag{2.42}$$

where p_f depends on k and the mapping $f \mapsto p_f$ is bounded from $L_m^2(B)$ into $\bar{W}$. Then

$$\langle p_f, \phi \rangle_{k_0} = \int_B \int f(\Delta_3 + k^2)\overline{\mathbf{S}^{-1}(k)\phi}\, d\mathbf{x} \tag{2.43}$$

and $\mathbf{S}^{-1}(k)$ is analytic since the isomorphism $\mathbf{S}(k)$ is also. Hence we have that $\overline{\mathbf{S}^{-1}(k)\phi}$ can be analytically continued off the real axis and from (2.43) the mapping $f \mapsto p_f$ is weakly analytic. Hence, the mapping is strongly analytic, i.e., $p_f = p_f(k)$ is an analytic function of k in a neighborhood of every point in R^+. We now define the analytic operator $\mathbf{P}_k : L_m^2(B) \to L_m^2(B)$ by

$$\mathbf{P}_k f = \frac{1}{m}(\Delta_3 + k^2)p_f. \tag{2.44}$$

Note that $\mathbf{P}_k f \in L_m^2(B)$ since

$$\begin{aligned} \int_B \int m|\mathbf{P}_k f|^2\, d\mathbf{x} &= \int_B \int \frac{1}{m}|\Delta_3 p_f + k^2 p_f|^2\, d\mathbf{x} \\ &= |p_f|_k^2 \\ &\leqq c\|f\|_{L_m^2(B)} \end{aligned} \tag{2.45}$$

for some constant c.

We now want to show that $\mathbf{P}_k$ is a projection operator for $k \in R^+$. To this end, let H be the vector space

$$H = \text{span}\,\{j_l(k|\mathbf{x}|)\, Y_l^m(\hat{\mathbf{x}}) : l = 0, 1, 2, \cdots, -l \leqq m \leqq l\}$$

where $j_l(k|\mathbf{x}|)$ is a spherical Bessel function and $Y_l^m(\hat{\mathbf{x}})$ is a spherical harmonic, and let $\bar{H}$ denote the closure of H in $L_m^2(B)$ and $H^\perp$ the orthogonal complement of H in $L_m^2(B)$. Then for $f \in H$ we have

$$\int_B \int \bar{f}(\Delta_3\phi + k^2\phi)\, d\mathbf{x} = \int_B \int \phi(\Delta_3\bar{f} + k^2\bar{f})\, d\mathbf{x} = 0 \tag{2.46}$$

for all $\phi \in \bar{W}$, and by a limiting process, (2.46) is also valid for $f \in \bar{H}$. Hence, since $\mathbf{S}$ is an isomorphism, $p_f = 0$ for $f \in \bar{H}$. For $f \in H^\perp$ we have

$$\begin{aligned} m(x)f(x) &= -(\Delta_3 + k^2)\int_B \int m(\mathbf{y})f(\mathbf{y})\Phi(\mathbf{x}, \mathbf{y})\, d\mathbf{y} \\ &= -(\Delta_3 + k^2)\mathbf{T}_k f \end{aligned} \tag{2.47}$$

and, by the addition formula for Bessel functions and the unique continuation property of solutions to the Helmholtz equation, $\mathbf{T}_k f$ is in $\bar{W}$. Hence from the definition of p_f we have

$$\begin{aligned} l_f(\mathbf{S}(k)\phi) &= \langle \mathbf{S}(k)\phi, p_f \rangle_{k_0} \\ &= \mathbf{B}(\phi, p_f; k), \end{aligned} \tag{2.48}$$

i.e., from (2.41) and (2.47),

$$\begin{aligned} &\int_B \int \frac{1}{m}(\Delta_3\overline{p_f} + k^2\overline{p_f})(\Delta_3\phi + k^2\phi)\, d\mathbf{x} \\ &\quad = -\int_B \int \frac{1}{m}(\Delta_3\overline{\mathbf{T}_k f} + k^2\overline{\mathbf{T}_k f})(\Delta_3\phi + k^2\phi)\, d\mathbf{x} \end{aligned} \tag{2.49}$$

for all $\phi \in \bar{W}$ and $k \in R^+$. Hence

$$p_f + \mathbf{T}_k f = 0 \tag{2.50}$$

for $f \in H^\perp$. Hence $\mathbf{P}_k f = f$ for $f \in H^\perp$, i.e., for $k \in R^+$, $\mathbf{P}_k$ is a projection operator (depending analytically on the parameter k).

We have shown above that $\mathbf{P}_k h = 0$ for $h \in \bar{H}$. In particular, if $\{v, w\}$ is a solution of (2.19) then $\mathbf{P}_k v = 0$. Furthermore, we see that $(1/k^2)(w - v) \in \bar{W}$ and for $k \in R^+$

$$\begin{aligned} l_w(\mathbf{S}(k)\phi) &= \int_B \int \bar{w}(\Delta_3 \phi + k^2 \phi)\, d\mathbf{x} \\ &= \int_B \int \frac{1}{m}(\Delta_3 + k^2)\left(\frac{\bar{w} - \bar{v}}{k^2}\right)(\Delta_3 + k^2)\phi\, d\mathbf{x} \\ &= \frac{1}{k^2}\langle \mathbf{S}(k)\phi, w - v\rangle_{k_0}, \end{aligned} \tag{2.51}$$

i.e., $p_w = (1/k^2)(w - v)$ and $\mathbf{P}_k w = (1/m)(\Delta_3 + k^2)((w - v)/k^2) = w$. If we now use Green's formula to rewrite (2.19) in the form

$$\begin{aligned} w(\mathbf{x}) - v(\mathbf{x}) &= -\int_B \int \Phi(\mathbf{x}, \mathbf{y})(\Delta_3 + k^2)(w(\mathbf{y}) - v(\mathbf{y}))\, d\mathbf{y} \\ &= -k^2(\mathbf{T}_k w)(\mathbf{x}) \end{aligned} \tag{2.52}$$

and apply the operator $\mathbf{P}_k$ to both sides of (2.52), we arrive at the operator equation

$$w + k^2 \mathbf{P}_k \mathbf{T}_k w = 0. \tag{2.53}$$

Since $\mathbf{P}_k$ is bounded, $\mathbf{P}_k \mathbf{T}_k$ is compact, and since $\mathbf{P}_k \mathbf{T}_k$ is an analytic function of k, we can conclude from the theory of analytic Fredholm operators that $(\mathbf{I} + k^2 \mathbf{P}_k \mathbf{T}_k)^{-1}$ is a meromorphic function of k (cf. [7]). Hence, from (2.53) $w = 0$, except possibly for a discrete set of values of k. From (2.19) this implies $v = 0$ and the theorem follows. □

The discrete set of values of $k > 0$ such that the set $\mathbf{F}$ is possibly not complete consists of those numbers $k > 0$ such that $\mathbf{I} + k^2 \mathbf{P}_k \mathbf{T}_k$ is not invertible. But $\mathbf{I} + k^2 \mathbf{P}_k \mathbf{T}_k$ not being invertible is equivalent to $\mathbf{I} + k^2 \mathbf{P}_k \mathbf{T}_k \mathbf{P}_k$ not being invertible, i.e., by the Fredholm alternative and the self-adjointness of $\mathbf{P}_k$, $\mathbf{I} + k^2 \mathbf{P}_k \mathbf{T}_k^* \mathbf{P}_k$ is not invertible. This is equivalent to saying that $\mathbf{I} + k^2 \mathbf{P}_k \mathbf{T}_k^*$ is not invertible. Note that in $L_m^2(B)$ we have that

$$\mathbf{T}_k^* v = \int_B \int \bar{\Phi}(\mathbf{x}, \mathbf{y}) m(\mathbf{y}) v(\mathbf{y})\, d\mathbf{y} \tag{2.54}$$

where $\bar{\Phi}(\mathbf{x}, \mathbf{y})$ denotes the compex conjugate of $\Phi(\mathbf{x}, y)$.

DEFINITION 1. Values of $k > 0$ such that $\mathbf{I} + k^2 \mathbf{P}_k \mathbf{T}_k^*$ is not invertible are called *transmission eigenvalues.*

As noted in the Introduction, transmission eigenvalues in general exist (cf. [3] for the special case of a spherically stratified medium). The above discussion shows that the transmission eigenvalues form a discrete set if (2.25) is valid.

3. Transmission eigenvalues and the interior transmission problem. In this section of our paper we will show how the transmission eigenvalues are related to the existence of nontrivial solutions to what we will call the homogeneous interior transmission problem. If k is not a transmission eigenvalue we will show that there exists a unique

solution to the interior transmission problem first discussed in [3]. The interior transmission problem we are about to consider arises in a natural way in connection with the inverse scattering problem of determining the index of refraction $n(\mathbf{x})$ from the far-field pattern $F(\hat{\mathbf{x}}; k, \hat{\alpha})$ for fixed k, $\hat{\alpha} \in \partial\Omega$, $\hat{\mathbf{x}} \in \partial\Omega$. This was first discussed in [3] and here we will merely outline the basic ideas. The "direct approach" for solving the inverse scattering problem is to determine $n(\mathbf{x})$ from (2.10) (recalling that $m(\mathbf{x}) = 1 - n(\mathbf{x})$) and the scattering problem (2.2)-(2.4). The "dual approach" introduced by Colton and Monk in [3] is to first determine $g \in L^2(\partial\Omega)$ such that

$$\int_{\partial\Omega} F(\hat{\mathbf{x}}; k, \hat{\alpha}) g(\hat{\mathbf{x}})\, ds(\hat{\mathbf{x}}) = 1 \tag{3.1}$$

for fixed k and $\hat{\alpha} \in \partial\Omega$. Then, if we define the Herglotz wave function v by (2.18) it can be shown as in Lemma 2 (see also [6]) that (3.1) is satisfied if and only if $\{v, w\}$ satisfies the *interior transmission problem*

$$\Delta_3 w + k^2 n(\mathbf{x}) w = 0 \qquad \text{in } \Omega_b, \tag{3.2}$$

$$\Delta_3 v + k^2 v = 0 \qquad \text{in } \Omega_b, \tag{3.3}$$

$$w(\mathbf{x}) - v(\mathbf{x}) = \frac{e^{-ikr}}{r} \qquad \text{on } \partial\Omega_b, \tag{3.4}$$

$$\frac{\partial w}{\partial r}(\mathbf{x}) - \frac{\partial v}{\partial r}(\mathbf{x}) = \frac{\partial}{\partial r}\frac{e^{-ikr}}{r} \qquad \text{on } \partial\Omega_b \tag{3.5}$$

where $\Omega_b = \{\mathbf{x}: |\mathbf{x}| < b\}$, $b > a$ (note that by unique continuation $v(\mathbf{x}) = w(\mathbf{x})$ for $\mathbf{x} \in \Omega_b \backslash B$). The "dual approach" for solving the inverse scattering problem is to now determine $n(\mathbf{x})$ from (3.1) and the interior transmission problem (3.2)-(3.5) [3]. From the point of view of applications, it suffices to consider only weak solutions of the interior transmission problem, since Theorem 3.3 of [3] shows that every weak solution, if it exists, can be approximated by a classical solution $\{v_0, w_0\}$ such that v_0 is a Herglotz wave function.

DEFINITION 2. The pair $\{v, w\}$ is said to be a *weak solution* of the interior transmission problem if

(a) $v \in \bar{H}$ (where H is defined in Theorem 1),

(b) $w \in L^2_m(B)$ satisfies $v = w + k^2 \mathbf{T}_k^* w$,

(c) For $\mathbf{x} \in \partial\Omega_b$, $k^2 \mathbf{T}_k^* w = -e^{-ikr}/r$.

As remarked after Lemma 2, the *homogeneous interior transmission problem* is to find a solution $\{v, w\}$ of (3.2), (3.3) such that

$$w(\mathbf{x}) - v(\mathbf{x}) = 0 \qquad \text{on } \partial\Omega_b, \tag{3.4'}$$

$$\frac{\partial w}{\partial r}(\mathbf{x}) - \frac{\partial v}{\partial r}(\mathbf{x}) = 0 \quad \text{on } \partial\Omega_b. \tag{3.5'}$$

The corresponding weak formulation is to find a pair $\{v, w\}$ such that (a) and (b) of Definition 2 are satisfied and for $x \in \partial\Omega_b$,

$$k^2 \mathbf{T}_k^* w = 0. \tag{3.6}$$

From (3.6) and the addition formula for Bessel functions we see that since $k > 0$, $w \in H^{\perp}$, and hence if $\{v, w\}$ is a weak solution of the homogeneous interior transmission problem then

$$0 = w + k^2 \mathbf{P}_k \mathbf{T}_k^* w, \tag{3.7}$$

i.e., for a nontrivial weak solution of the homogeneous interior transmission problem to exist for $k>0$, k must be a transmission eigenvalue. Conversely, if k is a transmission eigenvalue then there exists $w\in L^2_m(B)$, $w\neq 0$, such that (3.7) is valid. Then $w\in H^{\perp}$ and hence (3.6) is true by the addition formula for Bessel functions. Furthermore, setting $\mathbf{T}_k^* w = h+\mathbf{P}_k\mathbf{T}_k^* w$, $h\in\bar{H}$, we see that $v=w+k^2\mathbf{T}_k^* w$, where $v=k^2 h\in\bar{H}$, i.e., $\{v, w\}$ is a weak nontrivial solution of the homogeneous interior transmission problem.

In [3], the existence of a unique weak solution to the interior transmission problem is proved only for values of k sufficiently small. We can now do much better.

THEOREM 2. *Suppose $k>0$. Then if k is not a transmission eigenvalue there exists a unique weak solution to the interior transmission problem.*

Proof. If two weak solutions of the interior transmission problem exist then their difference satisfies the homogeneous interior transmission problem, and, since $k>0$ is not a transmission eigenvalue, from the discussion preceding this theorem the difference must be identically zero. Hence we have established uniqueness.

To prove existence, we introduce the vector space H_0 defined by

$$H_0=\text{span}\,\{j_l(k|\mathbf{x}|)\,Y_l^m(\hat{\mathbf{x}}):\ l=1,2,\cdots,\ -l\leqq m\leqq l\} \tag{3.8}$$

and let $\mathbf{P}_k^0: L^2_m(B)\to H_0^{\perp}$ be the projection operator from $L^2_m(B)$ onto the orthogonal complement $H_0^{\perp}$ of H_0. Suppose for $k>0$ we can find a constant c and function $w\in L^2_m(B)$ such that

$$c\mathbf{P}_k^0 j_0 = w+k^2\mathbf{P}_k\mathbf{T}_k^* w, \tag{3.9}$$

$$-k^2(j_0, w)=1 \tag{3.10}$$

where $j_0(k|\mathbf{x}|)$ is the spherical Bessel function of order zero. (We note that $\mathbf{P}_k^0 j_0\neq 0$.) Then $w\in H_0^{\perp}$ and setting $j_0=h_0+\mathbf{P}_k^0 j_0$, $h_0\in\bar{H}_0$, $\mathbf{T}_k^* w=h_1+\mathbf{P}_k\mathbf{T}_k^* w$, $h_1\in\bar{H}$, we see that

$$v=w+k^2\mathbf{T}_k^* w \tag{3.11}$$

where $v=c(j_0-h_0)+k^2 h_1\in\bar{H}$. From (3.10) and the fact that $w\in H_0^{\perp}$ we see from the addition formula for Bessel functions that part (c) of Definition 2 is satisfied, i.e., $\{v, w\}$ is a weak solution of the interior transmission problem. Hence to show existence it suffices to show the existence of a solution to (3.9), (3.10).

To solve (3.9), (3.10), it suffices to find a solution w_0 of

$$\mathbf{P}_k^0 j_0 = w_0+k^2\mathbf{P}_k\mathbf{T}_k^* w_0 \tag{3.12}$$

such that $(j_0, w_0)\neq 0$, since setting $\gamma=(j_0, w_0)$, $c=-1/k^2\gamma$, $w=-w_0/k^2\gamma$ gives a solution of (3.9), (3.10). But since k is not a transmission eigenvalue we know that there exists a unique solution of (3.12). Hence to complete the proof of the theorem it suffices to show that $(j_0, w_0)\neq 0$, where w_0 is the solution of (3.12). Suppose, on the contrary, that $(j_0, w_0)=0$. Then, since from (3.12) we have that $w_0\in H_0^{\perp}$, we can conclude that in fact $w_0\in H^{\perp}$. Then, since $\mathbf{P}_k\mathbf{P}_0^0 j_0=\mathbf{P}_k j_0=0$, we have from (3.12) that $w_0+k^2\mathbf{P}_k\mathbf{T}_k^* w_0=0$, which implies that $w_0=0$ since k is not a transmission eigenvalue. But from (3.12) this is a contradiction since $\mathbf{P}_k j_0\neq 0$. Hence $(j_0, w_0)\neq 0$ and the theorem is proved. □

In conclusion, we mention two possible topics for future research that we feel are particularly interesting. The first is to see if Theorems 1 and 2 are true when $n(\mathbf{x})$ can assume the value 1, i.e., the sound speed in the inhomogeneous medium can be both greater than and less than the sound speed of the host medium. (Note that if $n(\mathbf{x})$ is identically equal to 1 then, by the unique continuation principle for the Helmholtz equation, no solution exists to the interior transmission problem for any value of k.)

The second problem of interest to consider is the inverse spectral problem of determining the index of refraction $n(\mathbf{x})$ from the transmission eigenvalues. This is of particular interest for the inverse scattering problem since, as mentioned in the Introduction, the transmission eigenvalues can be numerically determined from the far-field data.

REFERENCES

[1] D. COLTON, *Dense sets and far field patterns for acoustic waves in an inhomogeneous medium*, Proc. Edinburgh Math. Soc. (2), 31 (1988), pp. 401–407.

[2] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[3] D. COLTON AND P. MONK, *The inverse scattering problem for time harmonic acoustic waves in an inhomogeneous medium*, Quart. J. Mech. Appl. Math., 41 (1988), pp. 97–125.

[4] D. COLTON AND L. PÄIVÄRINTA, *Far field patterns and the inverse scattering problem for electromagnetic waves in an inhomogeneous medium*, Math. Proc. Cambridge Philos. Soc., 103 (1988), pp. 561–575.

[5] P. HARTMAN AND C. WILCOX, *On solutions of the Helmholtz equation in exterior domains*, Math. Z., 75 (1961), pp. 228–255.

[6] A. KIRSCH, *Properties of far field operators in acoustic scattering*, Math. Methods Appl. Sci., to appear.

[7] M. REED AND B. SIMON, *Functional Analysis*, Academic Press, New York, 1972.

[8] P. WERNER, *Zur mathematischen Theorie akustischer Wellenfelder*, Arch. Rational Mech. Anal., 6 (1960), pp. 231–260.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1484–1489, November 1989

013

HÖLDER'S INEQUALITY FOR FUNCTIONS OF LINEARLY DEPENDENT ARGUMENTS*

FLORIN AVRAM† AND MURAD S. TAQQU‡

Abstract. Functions with values in L^p on a torus, in $L_p(-\infty,+\infty)$, or in l_p, $1 \leqq p \leqq \infty$ are considered. These functions are allowed to have linearly dependent arguments. A generalized Hölder inequality for products of such functions is established. The proofs involve the use of convex polytypes associated with polymatroids and the Riesz–Thorin interpolation theorem.

Key words. polymatroid, convexity, L_p spaces, power counting, Riesz–Thorin

AMS(MOS) subject classifications. primary 26A16; secondary 52A25

1. The generalized Hölder inequality.

A. Here we present the generalized Hölder inequality for products of functions of linearly dependent arguments obtained in [AB] for L_p functions, $1 \leqq p \leqq \infty$, on the torus. Corresponding results for other spaces such as $L_p(-\infty,+\infty)$ and l_p are also presented. These generalized Hölder inequalities can be used to obtain central and noncentral limit theorems for dependent random variables, e.g., time series with long-range dependence (see [T] and [A]). Long-range dependence is often encountered in nature, for example, in geophysical time series and in the context of critical phenomena in physics.

B. Let M be an $m \times n$ matrix, $\mathbf{x} = (x_1, \cdots, x_m)$, and let $l_1(\mathbf{x}), \cdots, l_n(\mathbf{x})$ be n linear transformations such that

$$(l_1(\mathbf{x}), \cdots, l_n(\mathbf{x})) = (x_1, \cdots, x_m)M.$$

Let $f_j, j = 1, \cdots, n$ be functions on $L_{p_j}(d\mu)$, $1 \leqq p_j \leqq \infty$.

We want to find conditions on $z_j = 1/p_j$, $j = 1, \cdots, n$, so that the generalized Hölder inequality

$$\text{(GH)} \qquad \left| \int \prod_{j=1}^{n} f_j(l_j(\mathbf{x})) \prod_{i=1}^{m} d\mu(x_i) \right| \leqq K \prod_{j=1}^{n} \|f_j\|_{p_j}$$

holds.

The key observation is that it is enough to find those points $\mathbf{z} = (z_1, \cdots, z_n)$ with coordinates z_i equal to zero or one, for which (GH) holds; then by the Riesz–Thorin interpolation theorem, (GH) will hold for the smallest convex set generated by these points.

C. We consider the following three cases:

(Ca) μ is Lebesgue measure, normalized to unity, on the torus and the matrix M has only integer elements.

(Cb) μ is a counting measure (i.e., $\int f(x)\, d\mu(x)$ means $\sum_{x=-\infty}^{+\infty} f(x)$). Moreover,

* Received by the editors January 13, 1988; accepted for publication (in revised form) January 20, 1989.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02215. This work was performed while the author was visiting Cornell University, Ithaca, New York 14853. This research was partially supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University, Ithaca, New York 14853.

‡ Department of Mathematics, Boston University, Boston, Massachusetts 02215. This research was supported by National Science Foundation grants ECS 8696-090 and DMS-88-05627, by Air Force Office of Scientific Research grant 89-0115 at Boston University, and by the Guggenheim Foundation.

the matrix M has only integer elements and all its nonsingular $m \times m$ minors have determinant ± 1.

(Cc) μ is Lebesgue measure and all the nonsingular $m \times m$ minors of M have determinants bounded below in absolute value by $1/K$.

D. Let A denote both a subset of columns of M and also the set of indices labeling the columns. Let $r(A)$ be the rank of the matrix with columns A. The subsets of M include $A = \varnothing$ whose rank is zero.

THEOREM. *Suppose, respectively, that conditions* (Ca), (Cb), *or* (Cc) *hold. Then the generalized Hölder inequality* (GH) *holds for any* $\mathbf{z} = (z_1, \cdots, z_n) \in [0, 1]^n$ *satisfying, respectively,*

(a1) $\sum_{j \in A} z_j \leqq r(A)$, *for all* A;

(b1) $\sum_{j \in A} z_j + r(A^c) \geqq r(M)$, *for all* A *and* $r(M) = m$;

(c1) $\sum_{j=1}^n z_j = r(M) = m$, *and either condition* (a1) *or* (b1).

The constant K *in* (GH) *equals* 1 *in cases* (Ca) *and* (Cb).

Remarks. (1) We can have $n < m$ in case (Ca) but not in cases (Cb) and (Cc). In fact, neither conditions (b1) or (c1) nor the generalized Hölder inequality (GH) are satisfied when $n < m$. (To verify that condition (b1) fails when $n < m$, take A to be all the columns of M. Then (b1) becomes $\sum_{j=1}^n z_j \geqq m$, having no solution $\mathbf{z} = (z_1, \cdots, z_n) \in [0, 1]^n$ when $n < m$.)

(2) The additional condition $\sum_{j=1}^n z_j = r(M)$ in (c1) says that (a1) (and (b1)) must be equalities when $A = M$.

In practice, it is not necessary to check that (a1) or (b1) hold for all A. It is enough to focus on flat and padded sets, which we now define.

Let $s(A)$ denote the *span* of A, i.e., the set of all linear combinations of columns in A that are contained in M. *Flats* are maximal dependent subsets, i.e., subsets of M that coincide with their span ($A = s(A)$). The set A is *padded* if any column $\mathbf{a}$ in A is also in $s(A \backslash \{\mathbf{a}\})$, i.e., if it is a linear combination of other columns of A. Note that a flat set is not necessarily a padded set (singletons are flats but not padded), nor is a padded set necessarily a flat set. For example, if M corresponds to the complete graph with 4 vertices, i.e.,

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

and if A is the subset consisting of the first four columns, then A is padded but not flat.

Because $z_i \geqq 0$, condition (a1) holds if it holds for all sets A that are flats. Condition (b1) holds if it holds for all sets A^c that are flats. In fact, because $0 \leqq z_i \leqq 1$, we have the following proposition.

PROPOSITION. *Conditions* (a1) *and* (b1) *in the theorem are, respectively, equivalent to*

(a1′) $r(A) - \sum_{j \in A} z_j \geqq 0$, *for all* A *flat and padded.*

(b1′) $r(M) - r(A) - \sum_{j \in A^c} z_j \leqq 0$, *for all* A *flat and padded and* $r(M) = m$.

Remarks. (1) The sets $\varnothing$ and M are flat and $\varnothing$ is padded. The set $A = \varnothing$ always satisfies (a1′) and the set $A = M$ always satisfies the inequality in (b1′).

(2) A is independent if $r(A) = |A|$. Any flat can be expressed as $s(A)$ where A is independent and, for such A, $r(s(A)) = r(A) = |A|$. Condition (a1′) is then equivalent to

$$|A| - \sum_{j \in s(A)} z_j \geqq 0 \quad \forall A \text{ independent.}$$

Such relations, known as power counting conditions in mathematical physics, ensure the convergence of the left-hand side of (GH) when the integrands are regularly varying functions (see, for example, [FT] and [TT]).

2. Examples.

(1) As an example of (a1'), consider the integral

$$J = \int_T \int_T f_1(x_1) f_2(x_2) f_3(x_1 + x_2) f_4(x_1 - x_2)\, dx_1\, dx_2$$

where T denotes the torus $[0, 1]$, so $f_j(x \pm 1) = f_j(x)$, $j = 1, \cdots, 4$. Here $m = 2$, $n = 4$, and

$$M = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \end{pmatrix}$$

has rank $r(M) = 2$. The flats consist of $\varnothing$, the single columns, and M. Only $\varnothing$ and M are flat and padded. Since (a1') always holds for $\varnothing$, it is sufficient to apply it to M. The theorem yields

$$|J| \leqq \|f_1\|_{1/z_1} \|f_2\|_{1/z_2} \|f_3\|_{1/z_3} \|f_4\|_{1/z_4}$$

for any $\mathbf{z} = (z_1, z_2, z_3, z_4) \in [0, 1]^4$ satisfying $z_1 + z_2 + z_3 + z_4 \leqq 2$, e.g., if $\mathbf{z} = (0, 1, \frac{1}{4}, \frac{1}{2})$, then

$$|J| \leqq \left(\sup_{0 \leqq x \leqq 1} |f_1(x)| \right) \left(\int_0^1 |f_2(x)|\, dx \right) \left(\int_0^1 f_3^4(x)\, dx \right)^{1/4} \left(\int_0^1 f_4^2(x)\, dx \right)^{1/2}.$$

(2) To illustrate (b1'), consider

$$S = \sum_{x_1 = -\infty}^{\infty} \sum_{x_2 = -\infty}^{\infty} f_1(x_1) f_2(x_2) f_3(x_1 + x_2) f_4(x_1 - x_2).$$

Since m, n, and M are as in Example (1), we have $r(M) = m$ and the only flat and padded sets are $\varnothing$ and M. Since it is sufficient to apply (b1') to $\varnothing$, the theorem yields $|S| \leqq \|f_1\|_{1/z_1} \|f_2\|_{1/z_2} \|f_3\|_{1/z_3} \|f_4\|_{1/z_4}$ for any $\mathbf{z} = (z_1, z_2, z_3, z_4) \in [0, 1]^4$ satisfying $z_1 + z_2 + z_3 + z_4 \geqq 2$, e.g., $|S| \leqq \prod_{j=1}^4 (\sum_{x=-\infty}^{+\infty} f_j^2(x))^{1/2}$.

(3) An application of (c1) yields

$$\left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(x_1) f_2(x_2) f_3(x_1 + x_2) f_4(x_1 - x_4)\, dx_1\, dx_2 \right| \leqq \prod_{j=1}^{4} \left(\int_{-\infty}^{+\infty} |f_j(x)|^{1/z_j}\, dx \right)^{z_j}$$

for any $\mathbf{z} = (z_1, z_2, z_3, z_4) \in [0, 1]^4$ satisfying $z_1 + z_2 + z_3 + z_4 = 2$.

(4) The usual Hölder inequality provides an extreme example of (c1), where $M = (1, \cdots, 1)$, $r(M) = m = 1$, and $r(A) = 1$, for all A. Then condition (c1) becomes $\sum_{j=1}^n z_j = 1$.

(5) The other extreme occurs when there are $n = m$ functions. If the linear transformations are of full rank ($r(M) = m$), then (c1) or a direct change of variable yields the trivial condition $z_j = 1$, $j = 1, \cdots, m$.

3. Proofs. Throughout, we consider the cases (a), (b), (c), characterized by the conditions (Ca), (Cb), (Cc), respectively.

LEMMA 1. *Let $\mathbf{z} = (z_1, \ldots, z_n)$ have all coordinates equal to zero or one, and let $A_{\mathbf{z}}$ denote the set of columns of the matrix M corresponding to those j's for which $z_j = 1$. Then* (GH) *holds if, respectively,*

(a2) $r(A_{\mathbf{z}}) = |A_{\mathbf{z}}|$ (*i.e., $A_{\mathbf{z}}$ is an "independent" set*);

(b2) $r(A_{\mathbf{z}}) = r(M) = m$ (*i.e., $A_{\mathbf{z}}$ is a "generating" set*);

(c2) $r(A_{\mathbf{z}}) = |A_{\mathbf{z}}| = r(M) = m$ (*i.e., $A_{\mathbf{z}}$ is a "basis"*).

The constant K in (GH) *equals* 1 *in cases* (a) *and* (b).

Proof. We replace $x_1, \cdots, x_m$ by a linear combination of new variables $u_1, \cdots, u_m$, as follows. In case (a), we have $r(A_{\mathbf{z}}) = |A_{\mathbf{z}}| \leqq m$. We let the $u_i = x_i$, $i = 1, \cdots, m - |A_{\mathbf{z}}|$ and set the remaining u_i's equal to the $l_j(\mathbf{x}), j \in A_{\mathbf{z}}$. In case (b) we let the u_i's be any independent subset of the $l_j(\mathbf{x}), j \in A_{\mathbf{z}}$. In case (c), we set $u_i = l_i(\mathbf{x})$, $i \in A_{\mathbf{z}}$.

In each of the cases, after switching to the new variables and bounding f_j by $|f_j|$, we can move the sup norms $\|f_j\|_{1/0}, j \notin A_{\mathbf{z}}$ outside the integrals. The conditions of the lemma ensure that what remains, namely $\prod_{j \in A_{\mathbf{z}}} |f_j|$, is integrable.

Note that in case (a), no constant K appears because the multiplicity of the linear map over the torus exactly cancels the Jacobian of the transformation. Note also that in the case (b), there may be more functions $|f_j|$ with $z_j = 1$ than variables u_j. But the multiple sum only increases if the arguments of these $|f_j|$ are changed to new arguments independent of u_i, $i = 1, \cdots, m$, and if these new arguments are summed. □

LEMMA 2. *The conditions* (a2), (b2), (c2) *and* (a1), (b1), (c1) *are, respectively, equivalent for points $\mathbf{z} = (z_1, \cdots, z_n)$ whose coordinates are all equal to zero or one.*

Proof. (a) $r(A_{\mathbf{z}}) = |A_{\mathbf{z}}|$ is equivalent to: for all $A \subset A_{\mathbf{z}}$, $r(A) = |A| = \sum_{j \in A} z_j$. For general A, write $A = A' \cup A''$, where $A' \subset A_{\mathbf{z}}$, $A'' \subset A_{\mathbf{z}}^c$. Then $\sum_{j \in A} z_j = \sum_{j \in A'} z_j = r(A') \leqq r(A' \cup A'') = r(A)$.

(b) One direction follows from

$$\begin{aligned} \sum_{j \in A} z_j + r(A^c) &= |A \cap A_{\mathbf{z}}| + r(A^c) \\ &\geqq r(A \cap A_{\mathbf{z}}) + r(A^c) \\ &\geqq r(A \cap A_{\mathbf{z}}) + r(A^c \cap A_{\mathbf{z}}) \\ &\geqq r(A_{\mathbf{z}}) \\ &= r(M) = m. \end{aligned}$$

To get the other direction, set $A^c = A_{\mathbf{z}}$.

(c) Suppose $\sum_{i=1}^n z_i = r(M) = m$. Then (a1) and (b1) are equivalent because when either one holds,

$$\sum_{j=1}^n z_j = \sum_{j \in A} z_j + \sum_{j \in A^c} z_j \leqq \sum_{j \in A} z_j + r(A^c).$$

So (c1)⇒(c2) by parts (a) and (b). To get (c2)⇒(c1) note that $|A_{\mathbf{z}}| = m \Leftrightarrow \sum_{j=1}^n z_j = m$, and use parts (a) and (b). □

Proof of the theorem. Consider the convex domain in $[0, 1]^n$ generated by $\mathbf{z} = (z_1, \cdots, z_n)$ satisfying condition (a1), (b1), (c1), respectively. The extreme points of the domain have coordinates all equal to zero or one.

Indeed, in case (a), condition (a1) characterizes the independent polytope of a polymatroid with the rank $r(\cdot)$ as submodular function [W, Thm. 18.3.1]. Edmonds' formula [W, Thm. 18.4.1] applies. The formula relates coordinates of the extreme points of the independence polytope to rank differences. Since all coordinates lie in [0, 1] and since the rank function is integer-valued, the coordinates of the extreme points can only be zero or one.

In case (b), let r^* be the rank function of the dual matroid [W, Thm. 2.1.2], i.e., $r^*(A)=|A|-r(M)+r(A^c)$. Condition (b1) can be rewritten $\sum_{j\in A}(1-z_j)\leqq |A|-r(M)+r(A^c)$, i.e.,

$$\sum_{j\in A}(1-z_j)\leqq r^*(A).$$

Edmonds' formula then applies to $1-z_j$.

Finally, in case (c), the extreme points for (c1) are a subset of the extreme points for (a1) and (b1).

Thus, by Lemmas 2 and 1, the generalized Hölder inequality (GH) holds for all $\mathbf{z}=(z_1,\ldots,z_n)$ that are extreme points of the convex domain characterized by (a1), (b1), (c1), respectively. To show that it holds for all $\mathbf{z}$ belonging to the convex domain use the Riesz-Thorin interpolation theorem [BL, Exercise 1.6.13, p. 18] with $z_j=1/p_j$ and $T=L_{p_1}\times\cdots\times L_{p_n}\to L_1$. □

Proof of the proposition. To prove (a1′)⇒(a1), we must show that $d(A)\equiv r(A)-\sum_{j\in A}z_j\geqq 0$ for all A. If A is padded and flat, then $d(A)\geqq 0$ by (a1′). If A is padded and not flat, then

$$d(A)\geqq r(A)-\sum_{j\in s(A)}z_j=r(s(A))-\sum_{j\in s(A)}z_j=d(s(A))\geqq 0 \tag{1}$$

by (a1′) since $s(A)$ is padded and flat. If A is not padded, there exists a column $\mathbf{a}_i\in A$ such that $\mathbf{a}_i\notin s(A\backslash\{\mathbf{a}_i\})$ are hence

$$d(A)=d(A\backslash\{\mathbf{a}_i\})+1-z_i\geqq d(A\backslash\{\mathbf{a}_i\}). \tag{2}$$

If $d(A\backslash\{\mathbf{a}_i\})\geqq 0$, we are done. If it is not, keep repeating the argument and use the fact that $d(\varnothing)\geqq 0$.

To prove (b1′)⇒(b1), we must show that $d'(A)\equiv r(M)-r(A)-\sum_{j\in A^c}z_j\geqq 0$ for all A. An argument similar to the preceding one applies. Merely replace (1) by

$$d'(A)\leqq r(M)-r(A)-\sum_{j\in(s(A))^c}z_j=r(M)-r(s(A))-\sum_{j\in(s(A))^c}z_j=d'(s(A))\leqq 0$$

and (2) by

$$d'(A)=d(A\backslash\{\mathbf{a}_i\})-1+z_i\leqq d(A\backslash\{\mathbf{a}_i\}).$$

We have $d(\varnothing)\leqq 0$ since $\varnothing$ is padded and flat. □

4. Extension. It is natural to expect that a generalized Hölder inequality holds for functions of several variables $f_j(x_1,\cdots,x_k)$ belonging to a tensor product space $L_{p_1}\otimes\cdots\otimes L_{p_k}$. There are many such tensor product spaces [LC]. The smallest one (corresponding to the greatest cross norm) is defined as follows: if f is a finite sum of products, then

$$\|f\|_{p_1,\cdots,p_k}=\inf\sum_{u=1}^{s}\prod_{v=1}^{k}\|f^{(u,v)}(x_v)\|_{p_v},$$

where the infimum is taken over all decompositions of $f(x_1,\cdots,x_k)$ in the form $\sum_{u=1}^{s}\prod_{v=1}^{k}f^{(u,v)}(x_v)$. The corresponding tensor space is obtained by completing the set of finite sums of products under this norm. A natural analogue of the theorem holds for such a space (e.g., see [AB, Thm. 1′]). It would be interesting to find out whether (GH) holds over larger L_p tensor product spaces (corresponding to smaller cross norms).

Acknowledgment. We thank Norma Terrin for useful discussions.

REFERENCES

[A] F. AVRAM, *Asymptotics of sums with dependent indices and convergence to the Gaussian distribution*, preprint, 1987.

[AB] F. AVRAM AND L. BROWN, *A generalized Hölder inequality and a generalized Szegö theorem*, 1987, Proc. Amer. Math. Soc., to appear.

[BL] J. BERGH AND J. LÖFSTROM, *Interpolation Spaces*, North-Holland, New York, 1976.

[FT] R. FOX AND M. TAQQU, *Central limit theorems for quadratic forms in random variables having long-range dependence*, Probab. Theory Rel. Fields, 74 (1987), pp. 213–240.

[LC] W. LIGHT AND E. CHENEY, *Approximation Theory in Tensor Product Spaces*, Lecture Notes in Mathematics, 1169, Springer-Verlag, Berlin, New York, 1980.

[TT] N. TERRIN AND M. S. TAQQU, *A noncentral limit theorem for quadratic forms of Gaussian stationary sequences*, 1988.

[T] M. S. TAQQU, *Toeplitz matrices and estimation of time series with long-range dependence*, in Proc. First World Congress of the Bernoulli Society, Tashkent (USSR), 1986, Yu. Prohorov and V. V. Sazonov, eds., Vol. 1, VNU Science Press. BV, Utrecht, The Netherlands, 1987, pp. 75–83.

[W] D. WELSH, *Matroid Theory*, Academic Press, London, 1976.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1490-1499, November 1989

014

FUCHSIAN SYSTEMS ASSOCIATED WITH THE $P^2(\mathbb{F}_2)$-ARRANGEMENT*

KOUICHI SAKURAI† AND MASAAKI YOSHIDA‡

Abstract. A family of Fuchsian systems of differential equations in two variables that interpolate Appell's F_1 and F_4 are constructed. Their geometric properties are also studied.

Key words. Fuchsian system, Appell's hypergeometric equations, orbifold, uniformization, Weyl group

AMS(MOS) subject classification. 33A35

0. Introduction. In this paper, we construct a family $E(s)$ of Fuchsian differential equations, depending on the two-dimensional parameter s, defined on the complex projective plane $M=\mathbb{C}P^2$ with regular singularities along

$$H: xyz(x-y)(y-z)(z-x)\{(x+y-z)^2-4xy\}=0$$

where $[x, y, z]$ is a system of homogeneous coordinates on M.

We call this arrangement H, of six lines and one conic, the $P^2(\mathbb{F}_2)$-arrangement, because the set of lines in the projective plane $P^2(\mathbb{F}_2)$ over the finite field $\mathbb{F}_2=\{0, 1\}$ consists of seven lines corresponding to the seven components of H. The arrangement H relates the Weyl group $W(F_4)$ of type F_4 as follows. The 24 mirrors of the reflections in the Coxeter group F_4 defines a hyperplane arrangement in $\mathbb{C}^4$. Passing to $\mathbb{C}P^3$, this arrangement defines a plane arrangement in $\mathbb{C}P^3$ called the $W(F_4)$-arrangement. The restriction $\tilde{H}$ of the $W(F_4)$-arrangement to any projective plane N in the arrangement consists of 13 lines in $\mathbb{C}P^2$ that can be given by the equation

$$\tilde{H}: XYZ(X^2-Y^2)(Y^2-Z^2)(Z^2-X^2)(X+Y+Z)(-X+Y+Z)(X-Y+Z)\cdot(X+Y-Z)=0$$

where $[X, Y, Z]$ is a system of homogeneous coordinates on N. The arrangement H is the image of $\tilde{H}$ under the map $\pi: N\to M$ given by

$$\pi: [X, Y, Z]\to[x, y, z]=[X^2, Y^2, Z^2].$$

The map π is the quotient map by the group $K(\simeq\mathbb{Z}_2+\mathbb{Z}_2)$ generated by $[X, Y, Z]\to[-X, Y, Z]$ and $[X, Y, Z]\to[X, -Y, Z]$ (see Fig. 1).

The two subarrangements

$$H': xyz(x-y)(y-z)(z-x)=0, \qquad H'': xyz\{(x+y-z)^2-4xy\}=0$$

of the arrangement H are well known as the singular loci of Appell's hypergeometric differential equations F_1 and F_4, respectively (see Fig. 2). Our equation interpolates the equation F_1 and the modified equation F_4' (see § 2) of F_4. More precisely, for some special values of the parameter s, our equation $E(s)$ turns out to be the equation F_1 and for some other special values of s, it turns out to be the equation F_4'. Moreover, the principal parts of the equations $E(s)$, after some normalization, are *linear* combinations of those of the equations F_1 and F_4'. We must say that it is a surprising result if we recall the nonlinearity of the integrability condition (see § 1). This mysterious phenomenon is also reported in [Yos2].

For some special values of the parameter s, the equation $E(s)$ happens to give the uniformizing equations (see § 3) of the hyperbolic orbifolds found by Hunt and

* Received by the editors March 7, 1988; accepted for publication February 8, 1989.

† Information Systems and Electronic Development Laboratory, Mitsubishi Electric Corporation, 5-1-1 Ofuna, Kamakura 247, Japan.

‡ Department of Mathematics, Faculty of Science, Kyushu University 33, Fukuoka 812, Japan.

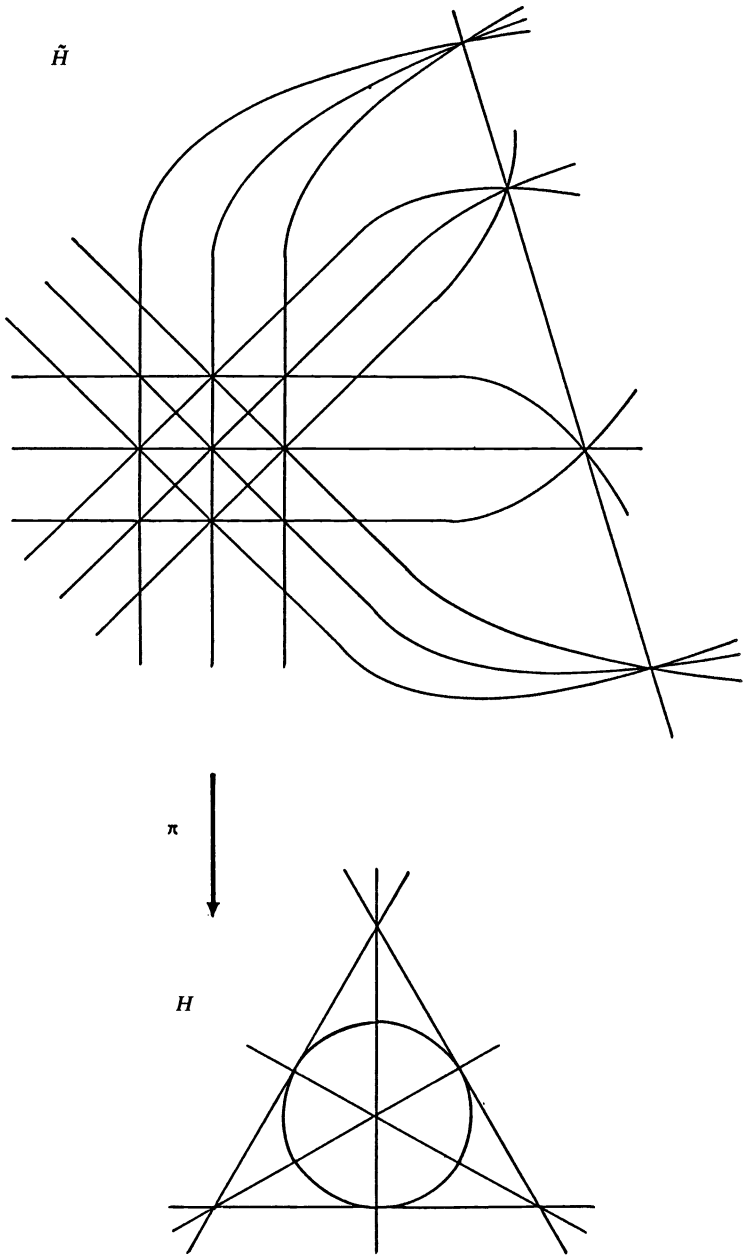


FIG. 1

Höfer [Hun], [Höf], where a hyperbolic orbifold is an orbifold whose universal uniformization is the complex unit ball $\mathbb{B}^2 = \{(z_1, z_2) \in \mathbb{C}^2 \mid |z_1|^2 + |z_2|^2 < 1\}$.

1. The main theorem. We consider a system in the form

$$\text{(E)} \qquad \frac{\partial^2 w}{\partial x_i \partial x_j} = \sum_{k=1}^{2} P_{ij}^k(x) \frac{\partial w}{\partial x_k} + P_{ij}^0(x) w, \qquad i, j = 1, 2$$

defined on $M = \mathbb{C}P^2$, where w is the unknown and $x = (x_1, x_2)$ is a system of inhomogeneous coordinates on M.

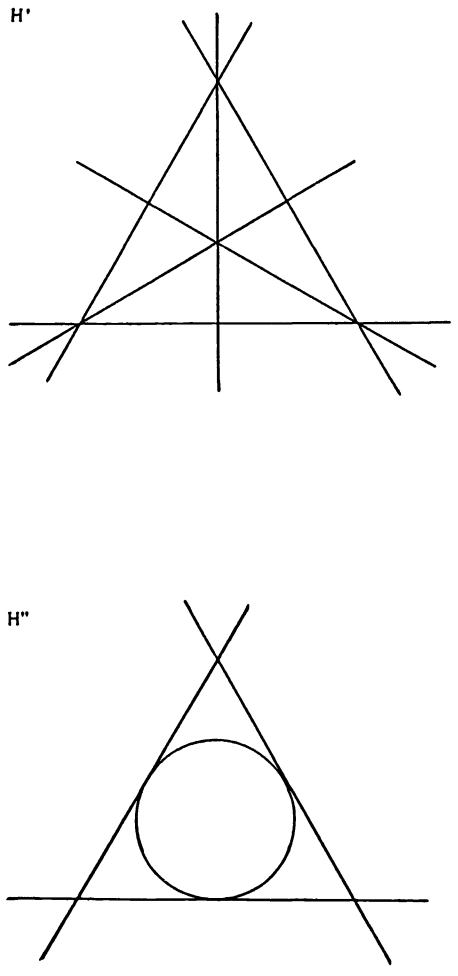


FIG. 2

DEFINITION. System (E) is said to be *in normal form* if

$$\text{(N)} \qquad \sum_{j=1}^{2} P_{ij}^{j}(x) = 0 \quad \text{for } i = 1, 2.$$

DEFINITION. System (E) is said to be *completely integrable* if (E) has three linearly independent solutions.

Any completely integrable system in the form (E) can be transformed into the uniquely determined system in normal form by replacing the unknown w by its product with a nonzero function of x. The consequent system is called the *normal form* of (E) [Yos1].

It is known [Yos3] that (E) in normal form is completely integrable if and only if the coefficients $\{P_{ij}^{k}\}$ satisfy the following equations:

$$P_{ij}^{k}(x) = P_{ji}^{k}(x), \qquad i, j = 1, 2, \quad k = 0, 1, 2,$$

$$P_{11}^{0}(x) = -\frac{\partial P_{11}^{1}(x)}{\partial x_1} - \frac{\partial P_{11}^{2}(x)}{\partial x_2} + 2\{P_{11}^{1}(x)\}^2 - 2P_{22}^{2}(x)P_{11}^{2}(x),$$

$$P^0_{12}(x)=\frac{\partial P^2_{22}(x)}{\partial x_1}+\frac{\partial P^1_{11}(x)}{\partial x_2}+P^1_{22}(x)P^2_{11}(x)-P^1_{11}(x)P^2_{22}(x),$$

$$P^0_{22}(x)=-\frac{\partial P^2_{22}(x)}{\partial x_2}-\frac{\partial P^1_{22}(x)}{\partial x_1}+2\{P^2_{22}(x)\}^2-2P^1_{11}(x)P^1_{22}(x),$$

$$(IC)_1 := -2\frac{\partial^2 P^1_{11}(x)}{\partial x_1\partial x_2}-\frac{\partial^2 P^2_{11}(x)}{\partial x_2^2}+6P^1_{11}(x)\frac{\partial P^1_{11}(x)}{\partial x_2}-3P^2_{11}(x)\frac{\partial P^2_{22}(x)}{\partial x_2}$$
$$-3P^2_{22}(x)\frac{\partial P^2_{11}(x)}{\partial x_2}+3P^1_{11}(x)\frac{\partial P^2_{22}(x)}{\partial x_1}-\frac{\partial^2 P^2_{22}(x)}{\partial x_1^2}$$
$$-2P^2_{11}(x)\frac{\partial P^2_{11}(x)}{\partial x_1}-P^1_{22}(x)\frac{\partial P^2_{11}(x)}{\partial x_1}=0,$$

$$(IC)_2 := -2\frac{\partial^2 P^2_{22}(x)}{\partial x_2\partial x_1}-\frac{\partial^2 P^1_{22}(x)}{\partial x_1^2}+6P^2_{22}(x)\frac{\partial P^2_{22}(x)}{\partial x_1}-3P^1_{22}(x)\frac{\partial P^1_{11}(x)}{\partial x_1}$$
$$-3P^1_{11}(x)\frac{\partial P^1_{22}(x)}{\partial x_1}+3P^2_{22}(x)\frac{\partial P^1_{11}(x)}{\partial x_2}-\frac{\partial^2 P^1_{11}(x)}{\partial x_2^2}$$
$$-2P^1_{22}(x)\frac{\partial P^2_{11}(x)}{\partial x_2}-P^2_{11}(x)\frac{\partial P^1_{22}(x)}{\partial x_2}=0.$$

LEMMA 1 [Yos3]. *Let $Q^k_{ij}(\xi)$ $(i,j,k=1,2)$ be the coefficients of the normal form of the transformed system of* (E) *by the coordinate change $\xi=\xi(x)$. Then we have*

$$Q^k_{ij}(\xi)=\sum_{l=1}^{2}\frac{\partial^2\xi_l}{\partial x_i\partial x_j}\frac{\partial x_k}{\partial \xi_l}-\frac{1}{3}\left(\delta^k_i\frac{\partial}{\partial x_j}+\delta^k_j\frac{\partial}{\partial x_i}\right)\log\left(\det\left(\frac{\partial\xi}{\partial x}\right)\right)$$
$$+\sum_{p,q,r=1}^{2}P^r_{pq}(x)\frac{\partial\xi_p}{\partial x_i}\frac{\partial\xi_q}{\partial x_j}\frac{\partial x_k}{\partial\xi_r}$$

where δ^k_j is the Kronecker symbol.

If $Q^k_{ij}(x)=P^k_{ij}(x)$ $(i,j,k=1,2)$, then the system (E) is said to be *invariant* under the transform $x\to\xi$.

DEFINITION. *A projective solution* of a completely integrable system (E) is the pair $z=(z_1,z_2)$ of ratios $z_i=w_i/w_0$ $(i=1,2)$ of three linearly independent solutions w_0, w_1, and w_2 of (E).

DEFINITION. A projective solution of (E) is said to be *ramifying at $O=(0,0)$ along $x_1=0$ with exponent α* if there exists a projective solution $z=(z_1,z_2)$, which is expressed as follows:

$$z_1(x)=x_1^\alpha v_1,\quad z_2(x)=v_2,\quad \det\left(\frac{\partial z}{\partial x}\right)=x_1^{\alpha-1}u$$

for some $\alpha\in\mathbb{C}$, where v_1, v_2, and u are holomorphic at O, not divisible by x_1.

LEMMA 2 [Yos2]. *If a projective solution of* (E) *in normal form is ramifying along $x_1=0$ with exponent α, then the coefficients P^k_{ij} of* (E) *have the following properties*:

(R) $P^2_{22}(x)$, $x_1P^2_{11}(x)$, $(1/x_1)\,P^1_{22}(x)$, *and* $P^1_{11}(x)-(\alpha-1/3x_1)$ *are holomorphic.*

DEFINITION. System (E) in the normal form is said to have *ramifying singularities* along $x_1=0$ with exponent α if the condition (R) holds.

To state the theorem we prepare some notation. Let $[x, y, z]$ be a system of homogeneous coordinates on M related to (x_1, x_2) by $x_1 = x/z$ and $x_2 = y/z$. Let $H_i (i = 1, \cdots, 7)$ denote the following curves:

$$H_1: \{x = 0\}, \quad H_2: \{y = 0\}, \quad H_3: \{z = 0\}, \quad H_4: \{x = y\},$$
$$H_5: \{y = z\}, \quad H_6: \{z = x\}, \quad H_7: \{(x+y-z)^2 - 4xy = 0\},$$

so we have $H = \bigcup_{i=1}^{7} H_i$. Let G be the transformation group on M generated by $[x, y, z] \to [z, x, y]$ and $[x, y, z] \to [y, x, z]$. Note that G is isomorphic to the symmetric group in three letters.

THEOREM. *For given complex numbers $s_i \neq 1$ $(i = 1, \cdots, 7)$, there is a completely integrable differential equation $E(s)$, in normal form, with ramifying singularities along H_i with exponent s_i if and only if*

$$s_1 = s_2 = s_3, \quad s_4 = s_5 = s_6, \quad 6s_1 - 3s_4 - 2s_7 + 2 = 0.$$

(In particular, $E(s)$ is G-invariant.) The four coefficients of $(E(s))$ are explicitly given as follows:

$$P^1_{11}(x_1, x_2) = \frac{S^1}{x_1} + \frac{S^4(x_1 - 2x_2 + 1)}{2(x_1 - 1)(x_1 - x_2)} + \frac{4S^7 x_2}{(x_1 + x_2 - 1)^2 - 4x_1 x_2},$$

$$P^2_{11}(x_1, x_2) = \frac{-3S^4 x_2(x_2 - 1)}{2x_1(x_1 - 1)(x_1 - x_2)} + \frac{4S^7 x_2(x_2 - 1)}{x_1\{(x_1 + x_2 - 1)^2 - 4x_1 x_2\}},$$

$$P^1_{22}(x_1, x_2) = P^2_{11}(x_2, x_1), \qquad P^2_{22}(x_1, x_2) = P^1_{11}(x_2, x_1),$$

where $S^i = \frac{1}{3}(s_i - 1)$.

Remark. Other coefficients P^k_{ij} of (E) are uniquely determined by the equalities (N) and by the assumption that (E) is completely integrable. Therefore, in the sequel, to describe the system (E) in normal form we give only the four coefficients P^1_{11}, P^2_{11}, P^2_{22}, and P^1_{22} of (E).

2. Relation between $E(s)$ and Appell's hypergeometric differential equations. Appell's hypergeometric equation $F_1(a, b, b', c)$ is a differential equation with regular singularities on $\bigcup_{i=1}^{6} H_i$, while it is nonsingular along H_7. The normal form of $F_1(a, b, b', c)$ is given by

$$P^1_{11}(a, b, b', c; x_1, x_2)$$
$$= \frac{1}{3} \frac{x_2(x_2 - 1)}{f_1} \{(c - b')x_2 + (2b - c)x_1 + (b' - (a + b + 1))x_1 x_2 + (a - b + 1)x_1^2\},$$

$$P^2_{11}(a, b, b', c; x_1, x_2) = \frac{(x_2(x_2 - 1))^2}{f_1} b,$$

$$P^2_{22}(a, b, b', c; x_1, x_2) = P^1_{11}(a, b', b, c; x_2, x_1),$$

$$P^1_{22}(a, b, b', c; x_1, x_2) = P^2_{11}(a, b', b, c; x_2, x_1),$$

where $f_1 = x_1 x_2 (x_1 - 1)(x_2 - 1)(x_1 - x_2)$.

PROPOSITION 1. *Equation $E(s)$ with $s_7 = 1$ coincides with the normal form of Appell's F_1 which is G-invariant. Precise correspondence between $E(s) = E(s_1, s_2, s_3)$ and $F_1(a, b, b', c)$ is given by*

$$E(s_1, s_4, 1) = F_1(-s_1 - s_4 + 1, -\tfrac{1}{2}(s_4 - 1), -\tfrac{1}{2}(s_4 - 1), -\tfrac{1}{2}(2s_1 + s_4 - 3)).$$

This identity is valid for all $s_1, s_4 \in \mathbb{C}$.

Appell's hypergeometric equation $F_4(a, b, c, c')$ is a differential equation with four linearly independent solutions and has regular singularities on $\bigcup_{i=1}^{3} H_i \cup H_7$, while it is nonsingular along $\bigcup_{i=4}^{6} H_i$. The solution space of $F_4(a, b, c, c')$ has three-dimensional invariant subspace if $b = c + c' + 1$ [Kat]. The corresponding equation is called the modified F_4 and denoted by $F_4'(a, b, c, c')$. The four coefficients of the normal form of $F_4'(a, b, c, c')$ are given by

$$P_{11}^1(a, b, c, c'; x_1, x_2) = -\frac{1}{3}\frac{1}{f_4}\{c(x_2-1)^2 + (a-(b+c+2c'-1))x_1 + (a+5b-c-2c'+1)x_1x_2 + (a-(b+2c'-1))x_1^2\},$$

$$P_{11}^2(a, b, c, c'; x_1, x_2) = \frac{x_2^2}{f_4}\{(a-b-c'+1)x_1 + (a+b-c'+1)(1-x_2)\},$$

$$P_{22}^2(a, b, c, c'; x_1, x_2) = P_{11}^1(a, b, c', c; x_2, x_1),$$

$$P_{22}^1(a, b, c, c'; x_1, x_2) = P_{11}^2(a, b, c', c; x_2, x_1),$$

where $f_4 = x_1x_2\{(x_1+x_2-1)^2 - 4x_1x_2\}$.

PROPOSITION 2. *Equation $E(s)$ with $s_4 = 1$ coincides with the normal form of Appell's modified F_4, which is G-invariant. Precise correspondence between $E(s)$ and $F_4'(a, b, c, c')$ is given by*

$$E(s_1, 1, 3s_1 - \tfrac{1}{2}) = F_4'(-3s_1+1, 2s_1+1, -s_1+1, -s_1+1).$$

3. Uniformizing equations of some hyperbolic orbifolds. We briefly recall the definitions of orbifolds and their uniformizations. Let X be a complex manifold, let S be a hypersurface of X, let $S = \bigcup_j S_j$ be its decomposition into irreducible components, and let b_j be either infinity or an integer called the weight attached to the corresponding S_j. The triple (X, S, b) is called an *orbifold* if for every point in $X - \bigcup \{S_j \mid b_j = \infty\}$ there is an open neighborhood U and a covering manifold that ramifies along $U \cap S$ with the given indices b. It is called *uniformizable* if there is a global covering manifold (called a *uniformization*) of X with the given ramification datum (S, b). If X is uniformizable, there exists a uniformization $\tilde{X}$ that is simply connected, called the *universal uniformization*. Let X be an orbifold and $\tilde{X}$ be its universal uniformization. The multivalued inverse map $X \to \tilde{X}$ of the projection $\tilde{X} \to X$ is called the *developing map*.

If the universal uniformization of an orbifold (X, S, b) is isomorphic to the complex ball (we call such an orbifold hyperbolic), there exists a unique Fuchsian differential equation in normal form such that its projective solution gives the developing map. The equation is called the *uniformizing equation* of the orbifold (X, S, b). For more details see [Yos2].

In his thesis, Hunt [Hun] has studied N-dimensional hyperbolic orbifolds. He has discovered a three-dimensional hyperbolic orbifold attached to the $W(\mathbb{F}_4)$-arrangement. Restricting to a plane in the $W(\mathbb{F}_4)$-arrangement, we have a two-dimensional hyperbolic orbifold attached to a line arrangement in $\mathbb{C}P^2$—the $P^2(\mathbb{F}_2)$-arrangement. Furthermore, Höfer [Höf] has shown that there are only four hyperbolic orbifolds over this arrangement. These orbifolds are given as follows. Consider the arrangement $\tilde{H}$ in N. Define the weight function b on N as in Table 1 (see Fig. 3).

COROLLARY. *Uniformizing differential equations of the above orbifolds are obtained by pulling back $E(s)$ under the map π, where the values of the parameter s are given as in Table 2.*

TABLE 1

Case	$(\cup_{i=1}^{3} \tilde{H}_i,$	$\cup_{i=4}^{6} \tilde{H}_i,$	$\tilde{H}_7,$	$p,$	$q,$	$r)$
1	$(\infty,$	2,	4,	−4,	2,	2)
2	(6,	2,	2,	−4,	6,	3)
3	(−6,	6,	2,	4,	2,	1)
4	(−3,	3,	∞,	∞,	1,	1)

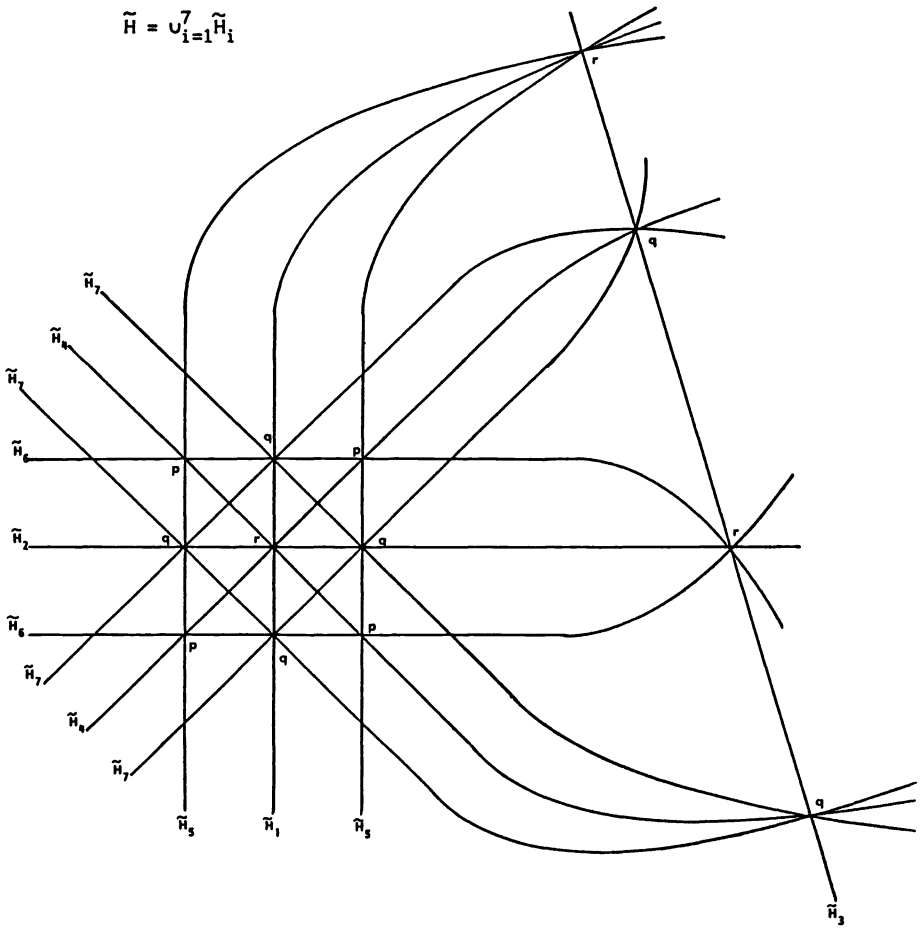


FIG. 3

TABLE 2

Case	$(s_1,$	$s_4,$	$s_7)$
1	(0,	$\frac{1}{2}$,	$\frac{1}{4}$)
2	$(\frac{1}{12},$	$\frac{1}{4}$,	$\frac{1}{4}$)
3	$(-\frac{1}{12},$	$\frac{1}{6}$,	$\frac{1}{4}$)
4	$(-\frac{1}{6},$	$\frac{1}{6}$,	0)

Remark. In the case where $s_7=1$, $E(s)$ reduces to Appell's F_1, which has been studied by Terada [Ter] and Deligne and Mostow [DM].

4. Proof of the results. Let the projective planes M, N, arrangements $\tilde{H}$, H, the group K, and the projection $\pi: N \to M$ be as above. The arrangement $\tilde{H}$ in N consists of 13 lines:

$$\tilde{H}_1=\{X=0\}, \quad \tilde{H}_2=\{Y=0\}, \quad \tilde{H}_3=\{Z=0\},$$
$$\tilde{H}_4=\{X^2-Y^2=0\}, \quad \tilde{H}_5=\{Y^2-Z^2=0\}, \quad \tilde{H}_6=\{Z^2-X^2=0\},$$
$$\tilde{H}_7=\{(X+Y+Z)(-X+Y+Z)(X-Y+Z)(X+Y-Z)=0\}.$$

Note that $\pi(\tilde{H}_i)=H_i$ $(i=1,\cdots,7)$ and $\pi(\tilde{H})=H$.

We construct a K-invariant differential equation $(\tilde{E})$ defined on N with ramifying singularities along $\tilde{H}_i$ with exponent t_i $(i=1,\cdots,7)$. We follow the method established in [Yos3].

LEMMA 3 [Yos3]. *If* (E) *has ramifying singularities along the line at infinity, then the total degree of the rational function* $P^k_{ij}(x)$ *is negative for* $i, j, k=1, 2$.

By Lemma 3, we can put

$$\tilde{P}^1_{11}=X_2(X_2^2-1)A/F, \quad \tilde{P}^2_{11}=X_2^2(X_2^2-1)^2B/F,$$
$$\tilde{P}^2_{22}=X_1(X_1^2-1)C/F, \quad \tilde{P}^1_{22}=X_1^2(X_1^2-1)^2D/F,$$

where

$$A=\sum_{i+j\leqq 8} a(i,j)X_1^iX_2^j, \quad B=\sum_{i+j\leqq 5} b(i,j)X_1^iX_2^j,$$
$$C=\sum_{i+j\leqq 8} c(i,j)X_1^iX_2^j, \quad D=\sum_{i+j\leqq 5} d(i,j)X_1^iX_2^j,$$
$$F=X_1X_2(X_1^2-1)(X_2^2-1)(X_1^2-X_2^2)(X_1+X_2+1)(-X_1+X_2+1)(X_1-X_2+1)\cdot(X_1+X_2-1).$$

The assumption that the system (E) is K-invariant says that

$$a(i,j)=c(i,j)=0 \quad \text{unless } i\equiv j\equiv 0 \bmod 2,$$
$$b(i,j)=0 \quad \text{unless } i+1\equiv j\equiv 0 \bmod 2,$$
$$d(i,j)=0 \quad \text{unless } i\equiv j+1\equiv 0 \bmod 2.$$

Applying Lemma 2, along every component of $\tilde{H}$, we obtain finitely many linear equations with unknowns $a(i,j),\cdots,d(i,j)$. By solving these, all the coefficients $a(i,j),\cdots,d(i,j)$ are expressed in terms of t_i $(i=1,\cdots,7)$:

$$\begin{aligned}A(X_1,X_2)=&-T^1(X_1^2-1)(X_1^2-X_2^2)\{(X_1^2+X_2^2-1)^2-4X_1^2X_2^2\}\\&-T^4X_1^2(X_1^2-2X_2^2+1)\{(X_1^2+X_2^2-1)^2-4X_1^2X_2^2\}\\&-8T^7X_1^2X_2^2(X_1^2-1)(X_1^2-X_2^2),\end{aligned}$$
$$B(X_1,X_2)=3T^4X_1\{(X_1^2+X_2^2-1)^2-4X_1^2X_2^2\}-8T^7X_1(X_1^2-1)(X_1^2-X_2^2),$$
$$C(X_1,X_2)=-A(X_2,X_1), \quad D(X_1,X_2)=-B(X_2,X_1),$$

where $T^i=\frac{1}{3}(t_i-1)$; moreover, these linear equations require that t_i's satisfy $T^1=T^2=T^3$, $T^4=T^5=T^6$, $3T^1-3T^4-2T^7=0$.

Now we study the integrability condition. The integrability condition of $\tilde{E}(t)$ is given by

$$(IC)_1=-64T^7(3T^1-3T^4-2T^7)X_1^2X_2^3(X_1^2-1)^2(X_2^2-1)^3(X_1^2-X_2^2)^2/F^2,$$
$$(IC)_2=-64T^7(3T^1-3T^4-2T^7)X_1^3X_2^2(X_1^2-1)^3(X_2^2-1)^2(X_1^2-X_2^2)^2/F^2.$$

Since the parameter has the relation such that $3T^1-3T^4-2T^7=0$, we have $(IC)_1=(IC)_2=0$.

Finally, we project $\tilde{E}(t)$ by the quotient map $\pi: M \to M/K \cong N$, where π is given by $(X_1, X_2) \to (x_1, x_2)=(X_1^2, X_2^2)$. By Lemma 1, coefficients P_{ij}^k of the normal form of $\pi(\tilde{E}(t))$ are given as follows:

$$P_{11}^1(x)=-\frac{1}{6x_1^2}+\frac{1}{2x_1}\tilde{P}_{11}^1(X), \qquad P_{22}^2(x)=-\frac{1}{6x_2^2}+\frac{1}{2x_2}\tilde{P}_{22}^2(X),$$

$$P_{11}^2(x)=\frac{x_2}{2x_1^2}\tilde{P}_{11}^2(X), \qquad P_{22}^1(x)=\frac{x_1}{2x_2^2}\tilde{P}_{22}^1(X).$$

Since π is branching along only the three lines $\bigcup_{i=1}^3 H_i$ with indices 2, there are the relations among the exponents such that $t_1=2s_1$, $t_4=s_4$, and $t_7=s_7$. Thus we obtain the differential equations we want. Easy calculations show the remaining claims of the theorem.

Propositions 1 and 2 are proved by straightforward computation, so we omit the detail. If $s_7=1$, the equations $(IC)_1=(IC)_2=0$ are satisfied. This proves the last statement of Proposition 1.

5. Linear structure of the set of solutions of the nonlinear differential equation (IC). The integrability condition

IC $$(IC)_1=(IC)_2=0$$

of (E) with (N) is a system of nonlinear differential equations with unknowns $\{P_{ij}^k\}=\{P_{11}^1=-P_{12}^2, P_{11}^2, P_{22}^1, P_{22}^2=-P_{12}^1\}$. There is a one-to-one correspondence between the set of solutions of **IC** and the set of completely integrable systems (E) in normal form. We have a great interest in rational solutions of **IC** of which corresponding systems (E) have transcendental solutions. The method used in § 4 is a practical one for finding such solutions.

Since the system **IC** is by no means linear, we cannot expect that linear combinations $\{t_1R_{ij}^k+t_2Q_{ij}^k\}(t_1, t_2 \in \mathbb{C})$ of two solutions $\{R_{ij}^k\}$ and $\{Q_{ij}^k\}$ of **IC** are also solutions of **IC**; indeed, it is not true in general. But sometimes miracles occur. Propositions 1 and 2 say that the coefficients $\{P_{ij}^k\}$ of the principal parts of $E(s)$ are linear combinations of the coefficients of the principal parts of the normal form of Appell's F_1 (with a special parameter) and those of the system F_4' (with a special parameter). In [Yos2], it is shown that linear combinations $\{t_1R_{ij}^k+t_2Q_{ij}^k\}$ of two solutions $\{R_{ij}^k\}$ and $\{Q_{ij}^k\}$ of **IC** are solutions of **IC**:

$\{R\}$
$$R_{11}^1(x, y)=3/x+81x^2y^3(2-x^3-y^3)/w,$$
$$R_{11}^2(x, y)=81xy(1+x^3-y^6-x^3y^3)/w,$$
$$R_{22}^2(x, y)=R_{11}^1(y, x), \qquad R_{22}^1(x, y)=R_{11}^2(y, x),$$

$\{Q\}$
$$Q_{11}^1(x, y)=3x^2(y^3-1)(1+x^3-2y^3)/2h,$$
$$Q_{11}^2(x, y)=-9xy(y^3-1)^2/2h,$$
$$Q_{22}^2(x, y)=Q_{11}^1(y, x), \qquad Q_{22}^1(x, y)=Q_{11}^2(y, x),$$

where

$$w=\prod_{a,b=0}^{2}(\omega^a x+\omega^b y+1)=(x^3+y^3+1)^3-27x^3y^3,$$
$$h=(x^3-1)(y^3-1)(x^3-y^3).$$

These two examples suggest the existence of some linear structure of the set of solutions of **IC** that is as yet veiled in mystery.

We conclude this paper by giving a useful system to test whether $t_1R_{ij}^k+t_2Q_{ij}^k$ $(t_1, t_2\in\mathbb{C})$ are solutions of **IC**.

PROPOSITION 3. *Let* $\{R_{ij}^k\}$ *and* $\{Q_{ij}^k\}$ *be solutions of* **IC**. *Then*:

(1) *For all* $t\in\mathbb{C}$, $\{tR_{ij}^k\}$ *are solutions of* **IC** *if and only if* $\{R_{ij}^k\}$ *satisfy the following conditions*:

$$2\frac{\partial^2 R_{11}^1}{\partial x_1\partial x_2}+\frac{\partial^2 R_{11}^2}{\partial x_2^2}+\frac{\partial^2 R_{22}^2}{\partial x_1^2}=0,$$

$$2\frac{\partial^2 R_{22}^2}{\partial x_2\partial x_1}+\frac{\partial^2 R_{22}^1}{\partial x_1^2}+\frac{\partial^2 R_{11}^1}{\partial x_2^2}=0.$$

(2) $\{R_{ij}^k+Q_{ij}^k\}$ *is a solution of* **IC** *if and only if* $(\{R_{ij}^k\},\{Q_{ij}^k\})$ *satisfy the following equations*:

$$\begin{aligned}
&6\left\{R_{11}^1\frac{\partial Q_{11}^1}{\partial x_2}+Q_{11}^1\frac{\partial R_{11}^1}{\partial x_2}\right\}\\
&\quad-3\left\{R_{11}^2\frac{\partial Q_{22}^2}{\partial x_2}+Q_{11}^2\frac{\partial R_{22}^2}{\partial x_2}-R_{22}^2\frac{\partial R_{11}^2}{\partial x_2}-Q_{22}^2\frac{\partial R_{11}^2}{\partial x_2}+R_{11}^1\frac{\partial Q_{22}^2}{\partial x_1}+Q_{11}^1\frac{\partial R_{22}^2}{\partial x_1}\right\}\\
&\quad-2\left\{R_{11}^2\frac{\partial Q_{22}^1}{\partial x_1}+Q_{11}^2\frac{\partial R_{22}^1}{\partial x_1}\right\}-\left\{R_{22}^1\frac{\partial Q_{11}^2}{\partial x_1}+Q_{22}^1\frac{\partial R_{11}^2}{\partial x_1}\right\}=0,\\
&6\left\{R_{22}^2\frac{\partial Q_{22}^2}{\partial x_1}+Q_{22}^2\frac{\partial R_{22}^2}{\partial x_1}\right\}\\
&\quad-3\left\{R_{22}^1\frac{\partial Q_{11}^1}{\partial x_1}+Q_{22}^1\frac{\partial R_{11}^1}{\partial x_1}-R_{11}^1\frac{\partial Q_{22}^1}{\partial x_1}-Q_{11}^1\frac{\partial R_{22}^1}{\partial x_1}+R_{22}^2\frac{\partial Q_{11}^1}{\partial x_2}+Q_{22}^2\frac{\partial R_{11}^1}{\partial x_2}\right\}\\
&\quad-2\left\{R_{22}^1\frac{\partial Q_{11}^2}{\partial x_2}+Q_{22}^1\frac{\partial R_{11}^2}{\partial x_2}\right\}-\left\{R_{11}^2\frac{\partial Q_{22}^1}{\partial x_2}+Q_{11}^2\frac{\partial R_{22}^1}{\partial x_2}\right\}=0.
\end{aligned}$$

Acknowledgment. The authors are grateful to Professor Mitsuo Kato who kindly informed us of the system F_4' and the usefulness of the system Reduce 3.2, which we used to carry out our complicated computation.

REFERENCES

[Bat] H. BATEMAN, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[DM] P. DELIGNE AND G. D. MOSTOW, *Monodromy of Hypergeometric Functions and Non-Lattice Integral Monodromy*, IHES Publication Mathematics, 63 (1986), pp. 5-106.

[Höf] TH. HÖFER, *Ballquotienten als verzweigte Überlegungen der projectiven Ebene*, Ph.D. dissertation, University of Bonn, Bonn, FRG, 1985.

[Hun] B. HUNT, *Coverings and Ball Quotients with Special Emphasis on the 3-Dimensional Case*, Bonner Math. Schriften, 174 (1986).

[Kat] MI. KATO, *A Pfaffian system of Appell's* F_4, Bull. of College Education, University of the Ryukyus, 33 (1988), pp. 331-334.

[Ter] T. TERADA, *Fonction hypergéométriques* F_1 *et fonctions automorphes* I, J. Math. Soc. Japan, 35 (1983), pp. 451-475.

[Yos1] M. YOSHIDA, *Canonical forms of some systems of linear partial differential equations*, Proc. Japan Acad. Ser. A., Math. Sci., 52 (1976), pp. 473-476.

[Yos2] ———, *Fuchsian Differential Equations*, Viehweg-Verlag, Wiesbaden, FRG, 1987.

[Yos3] ———, *Orbifold-uniformizing differential equations*, Math. Ann., 267 (1984), pp. 125-142.

SIAM J. MATH. ANAL.
Vol. 20, No. 6, pp. 1500-1513, November 1989

015

ON AN INTEGRAL TRANSFORM INVOLVING A CLASS OF MATHIEU FUNCTIONS*

D. NAYLOR†

Abstract. This paper constructs a formula of inversion for an integral transform whose kernel involves a type of Mathieu function of the third kind. The particular Mathieu function involved, $\psi = M_\nu^{(3)}(x+\frac{1}{2}i\pi)$, satisfies the equation $\psi_{xx} = (u^2+2h^2\cosh 2x)\psi$ together with the condition that $\psi(x,u)\to 0$ as $x\to\infty$. The transform in question can be used to generate solutions of the damped wave equation in infinite regions bounded by elliptic cylinders when a complex boundary condition is applied on the internal boundary.

Key words. integral transforms, Mathieu functions

AMS(MOS) subject classifications. 44A15, 33A55

1. Introduction. In a previous paper [2] the author considered the problem of determining the behavior for large values of u of a certain basic solution of the differential equation

$$y_{xx} = (u^2+2h^2\cosh 2x)y. \tag{1.1}$$

The solutions of this equation are related to the modified Mathieu functions $M_\nu^{(j)}(z)$, $j=1,2,3,4$, that satisfy

$$y_{zz} = (u^2-2h^2\cosh 2z)y. \tag{1.2}$$

Equation (1.1) transforms into (1.2) by means of the change of variable $z = x+\frac{1}{2}i\pi$ so that the functions $M_\nu^{(j)}(x+\frac{1}{2}i\pi)$ satisfy (1.1). The quantity ν, the characteristic exponent, used in the standard notation of the Mathieu functions is connected with the parameter u^2 by a complicated equation that [1] for large values of u assumes the form

$$u^2 = \nu^2 + O(h^4\nu^{-2}). \tag{1.3}$$

It is known [1] that

$$\begin{aligned} M_\nu^{(3)}(z) &= H_\nu^{(1)}(2h\cosh z)[1+O(\operatorname{sech} z)], \\ M_\nu^{(4)}(z) &= H_\nu^{(2)}(2h\cosh z)[1+O(\operatorname{sech} z)], \end{aligned} \tag{1.4}$$

as $\operatorname{Re}(z)\to\infty$ in any strip $|\operatorname{Im}(z)|\leqq$ constant, the parameter ν being held fixed. The functions $H_\nu^{(1)}$, $H_\nu^{(2)}$ denote the Hankel functions, the notation being that of Watson [8]. Since $\cosh(x+\frac{1}{2}i\pi) = i\sinh x$ and

$$\begin{aligned} H_\nu^{(1)}(ix) &= -\frac{2i}{\pi}\exp\left(-\frac{1}{2}i\nu\pi\right)K_\nu(x), \\ H_\nu^{(2)}(ix) &= \frac{2i}{\pi}\exp\left(-\frac{1}{2}i\nu\pi\right)K_\nu(x)+2\exp\left(\frac{1}{2}i\nu\pi\right)I_\nu(x), \end{aligned}$$

it follows that

$$\begin{aligned} M_\nu^{(3)}\left(x+\frac{1}{2}i\pi\right) &= -\frac{2i}{\pi}\exp\left(-\frac{1}{2}i\nu\pi\right)K_\nu(2h\sinh x)[1+O(\operatorname{cosech} x)], \\ M_\nu^{(4)}\left(x+\frac{1}{2}i\pi\right) &= \left[\frac{2i}{\pi}\exp\left(-\frac{1}{2}i\nu\pi\right)K_\nu(2h\sinh x)+2\exp\left(\frac{1}{2}i\nu\pi\right)I_\nu(2h\sinh x)\right] \\ &\quad\cdot[1+O(\operatorname{cosech} x)], \end{aligned}$$

* Received by the editors July 6, 1987; accepted for publication (in revised form) January 20, 1989.
† Department of Applied Mathematics, University of Western Ontario, London, Ontario, Canada N6A 5B9.

as $x \to \infty$. Finally, since [8, p. 202]

$$K_\nu(x) \sim (\pi/2x)^{1/2} e^{-x}, \qquad I_\nu(x) \sim (2\pi x)^{-1/2} e^{x},$$

as $x \to \infty$, we find that

$$\begin{aligned} M_\nu^{(3)}(x+\tfrac{1}{2}i\pi) &\sim -i(\pi h \sinh x)^{-1/2} \exp(-2h \sinh x - \tfrac{1}{2}i\nu\pi), \\ M_\nu^{(4)}(x+\tfrac{1}{2}i\pi) &\sim (\pi h \sinh x)^{-1/2} \exp(2h \sinh x + \tfrac{1}{2}i\nu\pi) \end{aligned} \tag{1.5}$$

as $x \to \infty$, for fixed ν. It follows from these formulas that there is essentially only one solution of (1.1) that remains bounded as $x \to +\infty$. This solution is the function $M_\nu^{(3)}(x+\frac{1}{2}i\pi)$, which for brevity we denote as $\psi(x, u)$. In [2] attention was concentrated on this solution, the object being to obtain sufficient information to investigate an integral transform having this function as kernel. Let $e^{-\lambda x} f(x) \in L^2(a, \infty)$; then the integral transform in question is defined by means of

$$F(u) = \int_a^\infty f(x)\psi(x, u)\, dx. \tag{1.6}$$

This transform is of use when it is required to construct solutions of the damped wave equation in regions exterior to cylinders of an elliptic cross section, in which case the radial eigenfunctions satisfy an equation such as (1.1), the quantity u^2 being an eigenvalue parameter, and it is the aim of the present paper to construct a formula of inversion for (1.6). The underlying expansion problem is that posed by the differential equation (1.1) on the interval $a \leqq x < \infty$ together with the condition that

$$ky(a) + y'(a) = 0 \tag{1.7}$$

where $k = k_1 + ik_2$ denotes a complex constant.

In the remainder of this section the asymptotic formulas derived in [2], giving the behavior of $\psi(x, u)$ as $u \to \infty$, are stated. In all of these formulas, x is supposed to be positive and fixed.

$$\begin{aligned} \psi(x, u) = 2\left(\frac{2}{\pi u}\right)^{1/2} &\exp\left(-\frac{1}{2}i\nu\pi + \frac{1}{2}iu\pi - \frac{1}{4}i\pi\right) \\ &\cdot \left\{\cosh\left[u \log\left(\frac{2u}{he}\right) - ux - \frac{1}{2}iu\pi - \frac{1}{4}i\pi\right] + O(u^{-1})\right\} \end{aligned} \tag{1.8}$$

as $u \to \infty$ in the sector $\frac{1}{2}\pi - \varepsilon \leqq \arg u \leqq \frac{1}{2}\pi$, and

$$\begin{aligned} \psi(x, u) = -2\left(\frac{2}{\pi u}\right)^{1/2} &\exp\left(-\frac{1}{2}i\nu\pi - \frac{1}{2}iu\pi + \frac{1}{4}i\pi\right) \\ &\cdot \left\{\cosh\left[u \log\left(\frac{2u}{he}\right) - ux + \frac{1}{2}iu\pi + \frac{1}{4}i\pi\right] + O(u^{-1})\right\} \end{aligned} \tag{1.9}$$

as $u \to \infty$ in the sector $-\frac{1}{2}\pi \leqq \arg u \leqq -\frac{1}{2}\pi + \varepsilon$. A simpler formula is valid as $u \to \infty$ in the sector $|\arg u| \leqq \frac{1}{2}\pi - \varepsilon$ where it can be shown that

$$\psi(x, u) = -i\left(\frac{2}{\pi u}\right)^{1/2} \exp\left[u \log\left(\frac{2u}{he}\right) - ux - \frac{1}{2}i\nu\pi\right][1 + O(u^{-1})]. \tag{1.10}$$

In fact, it will be shown in the next section of the paper that (1.10) also holds as $u \to \infty$ in the domain $\operatorname{Re}(u) \geqq c$ where c denotes a positive constant.

Formulas (1.8)-(1.10) were developed in [2] by applying formulas obtained by Pitts [4], who has also shown [5] that asymptotic formulas for the derivative $\psi_x(x, u)$ can be obtained from the corresponding formulas (1.8)-(1.10) by formal differentiation with respect to x so that, for example,

$$\psi_x(x, u) = -2\left(\frac{2u}{\pi}\right)^{1/2} \exp\left(-\frac{1}{2} i\nu\pi + \frac{1}{2} iu\pi - \frac{1}{4} i\pi\right) \cdot \left\{\sinh\left[u \log\left(\frac{2u}{he}\right) - ux - \frac{1}{2} iu\pi - \frac{1}{4} i\pi\right] + O(u^{-1})\right\} \tag{1.11}$$

as $u \to \infty$ in $\frac{1}{2}\pi - \varepsilon \leqq \arg u \leqq \frac{1}{2}\pi$.

2. The integral theorem. To construct a formula of inversion to be associated with the integral transform defined by (1.6), it is necessary to introduce a second basic solution $\phi(x, u)$ of (1.1). This solution will be defined as that generated by the initial values $\phi(a, u) = 1$, $\phi_x(a, u) = -k$ and so satisfies a condition like (1.7). The function $\phi(x, u)$ is necessarily a linear combination of $M_\nu^{(3)}$ and $M_\nu^{(4)}$, and it is easily verified that

$$\phi(x, u) = -\tfrac{1}{4} i\pi[g(u) M_\nu^{(4)}(x + \tfrac{1}{2} i\pi) - g_1(u) M_\nu^{(3)}(x + \tfrac{1}{2} i\pi)] \tag{2.1}$$

where

$$g(u) = kM_\nu^{(3)}(a + \tfrac{1}{2} i\pi) + M_\nu^{(3)'}(a + \tfrac{1}{2} i\pi), \tag{2.2}$$

$$g_1(u) = kM_\nu^{(4)}(a + \tfrac{1}{2} i\pi) + M_\nu^{(4)'}(a + \tfrac{1}{2} i\pi). \tag{2.3}$$

In this paper it is assumed that the constant k appearing in the boundary condition (1.7) is independent of u and in this case it can be shown [7], by writing the appropriate integral equation satisfied by $\phi(x, u)$, that this function is an entire function of the complex variable u and that for fixed values of x and large values of u in the halfplane $\operatorname{Re}(u) \geqq 0$,

$$\phi(x, u) = \cosh u(x - a) - ku^{-1} \sinh u(x - a) + O[u^{-2} e^{u(x-a)}]. \tag{2.4}$$

Before formulating the integral theorem, it is necessary to state certain results concerning the eigenvalues of the problem being investigated. These eigenvalues are the zeros p_n of the function $g(u)$ and it is assumed that they are arranged in ascending order of magnitude so that $|p_1| \leqq |p_2| \leqq |p_3| \cdots$. It is shown in the Appendix that if k is complex there are no real zeros and no purely imaginary zeros. If $\operatorname{Im}(k) > 0$ the zeros are located in the first and third quadrants of the complex u-plane, while if $\operatorname{Im}(k) < 0$ they are located in the second and fourth quadrants. Henceforth we will assume for definiteness that $\operatorname{Im}(k) > 0$. In this case, although no zero p_n can actually lie on the imaginary axis, it can be shown that those zeros of sufficiently large magnitude lie close to, and approach, that axis as their magnitudes tend to infinity, in the sense that $\operatorname{Re}(p_n) \to 0$ and $|\operatorname{Im}(p_n)| \to \infty$ as $n \to \infty$. It follows that all of the zeros can be positioned in a strip of finite width parallel to the imaginary axis of the complex u-plane.

THEOREM. *Let $f(x)$ be continuous for $x \geqq a > 0$ and suppose that $f(x)$ is of bounded variation in the neighborhood of x and that $e^{-\lambda x} f(x) \in L^2(a, \infty)$ where $\lambda \geqq 0$; then* (1.6) *implies*

$$f(x) = -\frac{1}{i\pi} \int_L \frac{u\phi(x, u) F(u)\, du}{g(u)} \tag{2.5}$$

where L denotes a line $\operatorname{Re}(u) = c$ parallel to the imaginary axis of the complex u-plane, $c > \max(\lambda, \frac{1}{2})$, and L is positioned so that all of the zeros of the function $g(u)$ lie to the left of it.

The method adopted to establish the above theorem will follow that used in [3], where an integral transform involving Bessel functions is discussed. This method stems from an inspection of equations (1.8)-(1.10) and (2.4), which show that when u is large the functions $\psi(x, u)$ and $\phi(x, u)$ behave for varying x like suitable combinations of the exponentials e^{ux} and e^{-ux}. By relating the Mathieu functions to the corresponding exponential or hyperbolic functions, the proof of the inversion formula (2.5) can be reduced to that of the Mellin inversion formula.

To verify (2.5), we let $L(R)$ denote the straight line in the complex u-plane drawn from $c-iR$ to $c+iR$, where c is the constant introduced in the theorem, and then form the equation

$$\int_{L(R)} \frac{u\phi(x, u)F(u)\,du}{g(u)} = \int_{L(R)} \frac{u\phi(x, u)\,du}{g(u)} \int_a^\infty f(t)\psi(t, u)\,dt. \tag{2.6}$$

This equation follows after inserting the expression (1.6) for $F(u)$. The order of integration of the repeated integral present on the right-hand side of (2.6) is now reversed. To justify this step it will be proved that, for finite values of R, the repeated integral in question exists when the integrand appearing therein is replaced by its modulus. With this purpose in mind we first apply the Schwarz inequality, revealing that

$$\left| \int_a^\infty f(t)\psi(t, u)\,dt \right| \leqq \|e^{-\lambda t}f(t)\| \cdot \|e^{\lambda t}\psi(t, u)\| \tag{2.7}$$

where $\|\cdot\|$ denotes the $L^2(a, \infty)$ norm so that

$$\|e^{\lambda t}\psi\| = \left(\int_a^\infty e^{2\lambda t}|\psi(t, u)|^2\,dt \right)^{1/2}. \tag{2.8}$$

A suitable bound on the value of this integral is obtained in the Appendix, where it is shown that

$$(u_1^2-\lambda^2)(u_2^2+\lambda^2) \int_a^\infty e^{2\lambda t}|\psi(t, u)|^2\,dt \leqq \frac{1}{4} u_1\, e^{2\lambda a}(|u\psi(a)|+|\psi'(a)|)^2 \tag{2.9}$$

where $u = u_1 + iu_2$ and $u_1 > \lambda$. It follows from (2.7) and (2.9) that the integral obtained from that on the right-hand side of (2.6), after taking the absolute values of all the terms in the integrand, does not exceed the quantity

$$\frac{\sqrt{c}\, e^{\lambda a}\|e^{-\lambda t}f\|}{2(c^2-\lambda^2)^{1/2}} \int_{-R}^{R} \left| \frac{u\phi(x, u)}{g(u)} \right| \frac{(|u\psi(a, u)|+|\psi_x(a, u)|)\,du_2}{(u_2^2+\lambda^2)^{1/2}}$$

since $u_1 = c$ on the path $L(R)$. This quantity is finite since the integrand is continuous on the path of integration and the latter is of finite extent. This permits the interchange of the order of integration in the integral on the right-hand side of (2.6) and leads to

$$\int_{L(R)} \frac{u\phi(x, u)F(u)\,du}{g(u)} = \int_a^\infty f(t)\,dt \int_{L(R)} \frac{u\phi(x, u)\psi(t, u)\,du}{g(u)}. \tag{2.10}$$

The next step is to determine the behavior of the integrand appearing on the right-hand side of (2.10) when the variable u is large and located in the halfplane $\mathrm{Re}\,(u) \geqq c$. First we note that, for large values of u in the stated halfplane, (1.8) and (1.9) both reduce to the simpler formula (1.10). That is to say, (1.10) holds as $u \to \infty$ in the halfplane $\mathrm{Re}\,(u) \geqq c$. To see this we write $u = re^{i\theta}$ and note that

$$u \log\left(\frac{2u}{he}\right) - ux - \frac{1}{2} iu\pi - \frac{1}{4} i\pi = A + iB \tag{2.11}$$

where

$$A = r\cos\theta\left[\log\left(\frac{2r}{he}\right) - x\right] + \left(\frac{1}{2}\pi - \theta\right) r\sin\theta,$$
$$B = r\sin\theta\left[\log\left(\frac{2r}{he}\right) - x\right] - \left(\frac{1}{2}\pi - \theta\right) r\cos\theta - \frac{1}{4}\pi. \tag{2.12}$$

Both terms on the right-hand side of (2.12) are nonnegative for r large and $0 \leqq \theta \leqq \frac{1}{2}\pi$. If u lies in the stated halfplane then $r\cos\theta \geqq c$, which is positive, and therefore A is not less than $c[\log(2r/he) - x]$, which is large and positive for sufficiently large r. Therefore $\cosh(A+iB) = \frac{1}{2}e^{A+iB} + O(u^{-c})$ where $c \geqq \frac{1}{2}$, and on making this change in (1.8) we find that the latter formula reduces to

$$\psi = -i\left(\frac{2}{\pi u}\right)^{1/2} \exp\left[u\log\left(\frac{2u}{he}\right) - ux - \frac{1}{2}i\nu\pi\right][1 + O(u^{-1})]. \tag{2.13}$$

Similarly, it can be shown that the real part of the argument appearing inside the hyperbolic function in (1.9) is also large and positive in the relevant part of the complex u-plane so that this function can be replaced by an exponential function. When this change is carried out we again obtain (2.13), which therefore holds as $u \to \infty$ throughout the region $\mathrm{Re}\,(u) \geqq c$. The asymptotic form of the function $g(u) = k\psi(a, u) + \psi_x(a, u)$ that also occurs in (2.10) can be deduced at once from the above formulas for ψ, from which it follows that

$$g(u) = i\left(\frac{2u}{\pi}\right)^{1/2} \exp\left[u\log\left(\frac{2u}{he}\right) - ua - \frac{1}{2}i\nu\pi\right][1 + O(u^{-1})] \tag{2.14}$$

as $u \to \infty$ in $\mathrm{Re}\,(u) \geqq c$. The asymptotic behavior of the function $\phi(x, u)$ also present in (2.10) is given by (2.4), and on combining this formula with (2.13) and (2.14) we find that

$$\frac{u\phi(x, u)\psi(t, u)}{g(u)} = -\frac{1}{2}e^{u(x-t)} - \frac{1}{2}e^{u(2a-x-t)} + h(u, x, t) \tag{2.15}$$

where $h(u, x, t) = O[u^{-1}e^{u(x-t)}]$ as $u \to \infty$ in $\mathrm{Re}\,(u) \geqq c$, uniformly on finite x and t intervals. Since $\phi(x, u)$, $\psi(t, u)$, $g(u)$ are analytic functions of u, the function $h(u, x, t)$ is also analytic except for simple poles at the zeros p_n of $g(u)$.

On inserting (2.15) into the integral on the right-hand side of (2.10) we obtain

$$\begin{aligned}\int_{L(R)} \frac{u\phi(x, u)F(u)\,du}{g(u)} = &-\frac{1}{2}\int_{L(R)} e^{ux}\,du \int_a^\infty e^{-ut}f(t)\,dt \\ &-\frac{1}{2}\int_{L(R)} e^{u(2a-x)}\,du \int_a^\infty e^{-ut}f(t)\,dt \\ &+\int_a^\infty f(t)\,dt \int_{L(R)} h(u, x, t)\,du\end{aligned} \tag{2.16}$$

after inverting the order of integration in the first and second repeated integrals on the right-hand side of (2.16). This process is again justified for finite values of R by absolute convergence since on writing $e^{-ct}f = e^{-(c-\lambda)t}\cdot e^{-\lambda t}f$, it is found after using the Schwarz inequality that $e^{-\lambda t}f \in L^2(a, \infty)$ implies $e^{-ct}f \in L(a, \infty)$.

Now by the Mellin inversion theorem [6], since $x > a$ we find that

$$\lim_{R\to\infty} \int_{c-iR}^{c+iR} e^{ux}\,du \int_a^\infty e^{-ut}f(t)\,dt = 2i\pi f(x) \tag{2.17}$$

while by the same theorem, the second term on the right-hand side of (2.16) tends to zero as $R \to \infty$ since $2a - x < a$ therein. Hence, on proceeding to the limit as $R \to \infty$ in (2.16), we obtain

$$\lim_{R\to\infty} \int_{L(R)} \frac{u\phi(x,u)F(u)\,du}{g(u)} = -i\pi f(x) + \lim_{R\to\infty} \int_a^\infty f(t)\,dt \int_{L(R)} h(u,x,t)\,du. \tag{2.18}$$

To complete the proof of (2.5) of the theorem it is necessary to verify that

$$\lim_{R\to\infty} \int_a^\infty f(t)\,dt \int_{L(R)} h(u,x,t)\,du = 0. \tag{2.19}$$

To obtain this result it will first be shown that

$$\int_{L(R)} [h(u,x,t) - h(u,t,x)]\,du = O(R^{-1}) \tag{2.20}$$

as $R \to \infty$, uniformly for any bounded interval of values of t. This result will be established with the aid of Cauchy's theorem and for this purpose we need some properties of the function $h(u,x,t) - h(u,t,x)$, which by (2.15) is given by

$$h(u,x,t) - h(u,t,x) = \sinh u(x-t) + uj(u,x,t) \tag{2.21}$$

where

$$j(u,x,t) = [\phi(x,u)\psi(t,u) - \phi(t,u)\psi(x,u)]/g(u).$$

It is shown in the argument that follows that the function on the left-hand side of (2.21) is an odd function of u that is $O(u^{-1}\, e^{u|x-t|})$ as $u \to \infty$ throughout the complex u-plane. The function $j(u,x,t)$, regarded as a function of x, satisfies the basic equation (1.1) and automatically vanishes at the value $x = t$. Furthermore, the derivative of this function with respect to x, also evaluated at $x = t$, equals $-W(\phi,\psi)/g(u)$, where $W(\phi,\psi)$ denotes the Wronskian. Since this Wronskian equals $k\psi(a,u) + \psi_x(a,u)$, its value at $x = a$, and this by (2.2) equals $g(u)$, it follows that $j(u,x,t)$ is the solution of (1.1) defined by the initial values $j(u,t,t) = 0$, $j_x(u,t,t) = -1$. Therefore $j(u,x,t)$ is an entire function of u^2, that is, an even entire function of u and, in addition [7, p. 10],

$$j(u,x,t) = -u^{-1} \sinh u(x-t) + O(u^{-2}\, e^{u|x-t|})$$

as $u \to \infty$ throughout the u-plane, uniformly for any bounded intervals of values of x and t. It follows from (2.21) that the function $h(u,x,t) - h(u,t,x)$ is, as stated, an odd function of u that is entire and that is $O(u^{-1}\, e^{u|x-t|})$ as $u \to \infty$ uniformly on any bounded interval of values of t.

Equation (2.20) may now be obtained by integrating the function $h(u,x,t) - h(u,t,x)$ around the boundary of the rectangle $0 \leqq \mathrm{Re}\,(u) \leqq c$, $-R \leqq \mathrm{Im}\,(u) \leqq R$. Since this function is odd in u the contribution from the line joining the points $\pm iR$ vanishes while the stated asymptotic bound on this function shows that the contributions from the top and bottom sides of the rectangle are each $O(R^{-1})$. This leaves the contribution from the remaining side $L(R)$ as $O(R^{-1})$, as stated in (2.20), since by Cauchy's theorem the sum of the integrals around all four sides is zero.

We may now return to the proof of (2.19). The integral present in the latter equation will be expressed as the sum of two integrals, obtained by dividing the t-interval into the parts (a,x) and (x,∞), and it will be shown that both of these integrals tend to zero as $R \to \infty$.

We consider first the integral

$$\int_a^x f(t)\,dt \int_{L(R)} h(u, x, t)\,du. \tag{2.22}$$

According to the relation (2.20) the expression (2.22) may be rewritten as the integral

$$\int_a^x f(t)\,dt \int_{L(R)} h(u, t, x)\,du + O(R^{-1}). \tag{2.23}$$

An estimate of the inner integral, the one along the path $L(R)$, occurring in this expression can be obtained by deforming the path onto the semicircle drawn to the right of $L(R)$ and having this line as diameter. On this semicircle

$$h(u, t, x) = O[R^{-1}\, e^{-(x-t)R\cos\theta}]$$

so that

$$\left|\int_{L(R)} h(u, t, x)\,du\right| \leqq \int_{-\pi/2}^{\pi/2} e^{-(x-t)R\cos\theta}\,d\theta$$

$$\leqq \begin{cases} \pi R^{-1}(x-t)^{-1}, & a \leqq t < x, \\ \pi, & a \leqq t \leqq x. \end{cases} \tag{2.24}$$

In view of the fact that the first of the two bounds on the right-hand side of (2.24) is not suitable in the vicinity of the value $t = x$, we subdivide the t-integration in (2.23) into the parts $(a, x - R^{-1/2})$ and $(x - R^{-1/2}, x)$, applying the first bound to the first subinterval and the second bound to the second subinterval. It is then found that the modulus of (2.23) is not greater than the quantity

$$\pi R^{-1} \int_a^{x-R^{-1/2}} |f(t)|(x-t)^{-1}\,dt + \pi \int_{x-R^{-1/2}}^{x} |f(t)|\,dt + O(R^{-1}). \tag{2.25}$$

Both integrals here are $O(R^{-1/2})$ since $(x-t)^{-1} \leqq R^{1/2}$ in the first integral and the domain of integration in the second integral is of length $R^{-1/2}$. Therefore all three terms appearing in (2.25) tend to zero as $R \to \infty$.

We consider next the integral

$$\int_x^\infty f(t)\,dt \int_{L(R)} h(u, x, t)\,du. \tag{2.26}$$

To treat this integral, the interval (x, ∞) is decomposed into the parts (x, x_0) and (x_0, ∞), where x_0 is chosen greater than x and large enough to ensure that

$$\left(\int_{x_0}^\infty e^{-2\lambda t}|f(t)|^2\,dt\right) < \varepsilon \tag{2.27}$$

where ε is arbitrarily small and positive.

The contribution of the interval (x, x_0) to the value of the integral (2.26) is the integral

$$\int_x^{x_0} f(t)\,dt \int_{L(R)} h(u, x, t)\,du. \tag{2.28}$$

Since $h(u, x, t) = O[u^{-1}\, e^{u(x-t)}]$, where now $t \geqq x$, and this tends to zero as $u \to \infty$ in the region on the right-hand side of $\mathrm{Re}\,(u) = c$, it can be seen on following an argument similar to that used to discuss the integral (2.22) that the integral (2.28) also tends to zero as $R \to \infty$.

To complete the analysis of (2.26) it will now be proved that

$$\left|\int_{x_0}^{\infty} f(t)\,dt \int_{L(R)} h(u, x, t)\,du\right| \leqq M\varepsilon \tag{2.29}$$

where M is a constant, and R large enough. In view of the definition (2.15) of the function $h(u, x, t)$, the integral present in (2.29) is equal to the sum

$$\begin{aligned}&\frac{1}{2}\int_{x_0}^{\infty} f(t)\,dt \int_{L(R)} e^{u(x-t)}\,du+\frac{1}{2}\int_{x_0}^{\infty} f(t)\,dt \int_{L(R)} e^{u(2a-x-t)}\,du\\ &\quad+\int_{x_0}^{\infty} f(t)\,dt \int_{L(R)} \frac{u\phi(x, u)\psi(t, u)\,du}{g(u)}.\end{aligned} \tag{2.30}$$

The first two repeated integrals appearing in (2.30) are similar to those on the right-hand side of (2.16) and vanish as $R\to\infty$, as is seen on reversing the order of integration and applying the Mellin inversion theorem, since x and $2a-x$ both lie outside the interval of integration (x_0, ∞) in (2.30).

The third double integral present in (2.30) may be shown to be $O(\varepsilon)$ by deforming the path $L(R)$ onto the rays $W(R)$ and the circular arcs $C(R)$ depicted in Fig. 1. The rays are defined by $u = c + s\,e^{\pm i\theta_0}$ where s varies from zero to R and θ_0 is an acute angle, and the arcs connecting the extremities of $L(R)$ and $W(R)$ are of radius R. On replacing $L(R)$ by $W(R)$ and $C(R)$, the double integral in question is replaced by

$$\int_{W(R)} \frac{u\phi(x, u)\,du}{g(u)} \int_{x_0}^{\infty} f(t)\psi(t, u)\,dt+\int_{C(R)} \frac{u\phi(x, u)\,du}{g(u)} \int_{x_0}^{\infty} f(t)\psi(t, u)\,dt \tag{2.31}$$

after reversing the orders of integration in both repeated integrals, a process that is permissible since it will be apparent in the course of the argument that follows that both integrals are absolutely convergent. The inner integrals present in (2.31) are estimated by means of the Schwarz inequality, which, by analogy with (2.7), shows in view of (A9) with λ replaced by 2λ therein that

$$\int_{x_0}^{\infty} |f(t)\psi(t, u)|\,dt \leqq \frac{\varepsilon(u_1)^{1/2}\,e^{\lambda a}|u\psi(x_0, u)-\psi_x(x_0, u)|}{4(u_1^2-\lambda^2)^{1/2}(u_2^2+\lambda^2)^{1/2}}.$$

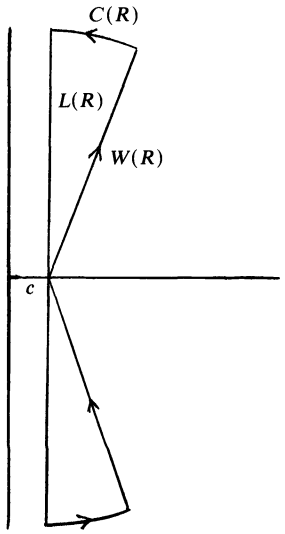


FIG. 1. *Complex u-plane.*

On using this result together with (2.4), (2.13), and (2.14) to estimate the values of ϕ, ψ, $g(u)$ for large values of u, it is found that

$$\left|\frac{u\phi(x,u)}{g(u)}\int_{x_0}^{\infty}|f(t)\psi(t,u)|\,dt\right| \leqq \frac{\varepsilon u_1^{1/2}\,e^{\lambda a}|u\,e^{-u(x_0-x)}|}{2(u_1^2-\lambda^2)^{1/2}(u_2^2+\lambda^2)^{1/2}}$$

$$\leqq \frac{\varepsilon\exp(\lambda a-(x_0-x)R\cos\theta)}{2(c-\lambda)^{1/2}\sin\theta} \tag{2.32}$$

since $u_1^2-\lambda^2=(u_1-\lambda)(u_1+\lambda)\geqq(c-\lambda)u_1$.

The expression (2.32) is integrable along the path $W(R)$ so that the first repeated integral in (2.31) exists as $R\to\infty$ and is clearly $O(\varepsilon)$. The use of the bound (2.32) on the arcs $C(R)$ shows that the second repeated integral in (2.31) does not exceed the quantity

$$\frac{\varepsilon R\,e^{\lambda a}}{2(c-\lambda)^{1/2}\sin\theta_0}\int_{\theta_0}^{\pi/2}e^{-(x_0-x)R\cos\theta}\,d\theta.$$

It is readily shown that the integral appearing in this expression is $O(R^{-1})$ so that the whole expression is $O(\varepsilon)$ for arbitrarily large values of R. It follows that both of the repeated integrals in (2.31) are $O(\varepsilon)$ and so the same is true of the third such integral in (2.30). Since the other two integrals in (2.30) tend to zero as $R\to\infty$, the result (2.29) follows for large enough values of R. Since (2.22) and (2.28) tend to zero as $R\to\infty$, the result (2.19) follows on adding all three integrals (2.22), (2.28), and (2.29). Finally, as a consequence of (2.19), (2.18) reduces to the inversion formula stated in the theorem.

3. The eigenfunction expansion. An explicit expansion in eigenfunctions can be deduced from (2.5) by deforming the path of integration onto the imaginary axis and taking the residues at the poles traversed. Since the functions $\phi(x,u)$, $\psi(x,u)$, and therefore $F(u)$, $g(u)$, are even functions of u, the integral taken along the entire imaginary axis vanishes, leaving the contributions from the residues at the zeros of $g(u)$. Since at such a zero the function $\phi(x,u)$ defined by (2.1) reduces to $\frac{1}{4}i\pi g_1(u)\psi(x,u)$, we find on calculating the residues the expansion

$$f(x)=-\frac{1}{2}i\pi\sum_{n=1}^{\infty}\frac{p_n g_1(p_n)\psi(x,p_n)F(p_n)}{g'(p_n)} \tag{3.1}$$

where the summation extends over all of the zeros p_n of $g(u)$ that lie in the first quadrant of the complex u-plane.

To justify the procedure it will be assumed that $f\in L^2(a,\infty)$ and we revert to (2.5) rewritten in the form

$$f(x)=-\frac{1}{i\pi}\lim_{R\to\infty}\int_{c-iR}^{c+iR}\frac{u\phi(x,u)F(u)\,du}{g(u)}. \tag{3.2}$$

We now apply Cauchy's theorem to the rectangle $0\leqq\operatorname{Re}(u)\leqq c$, $-R\leqq\operatorname{Im}(u)\leqq R$, where R is chosen so that the top side of the rectangle passes between consecutive poles of the integrand. To achieve this the quantity R will be allowed to tend to infinity through the sequence of values R_n defined by $R_n\log(2R_n/he)-aR_n=n$, where n is an integer that tends to infinity. With this choice the sides of the rectangle will, by (A15), avoid the poles. In addition, we note that the quantity B, introduced in (2.11) with x set equal to a therein, will be such that $B\sim(n+\frac{1}{4})\pi$ as $n\to\infty$. We now verify

that the integrals along the top and bottom sides vanish in the limit as $R \to \infty$. The integral along the top side is equal to

$$J = \int_0^c \frac{u\phi(x, u)\, du_1}{g(u)} \int_a^\infty f(t)\psi(t, u)\, dt. \tag{3.3}$$

We again divide the integration with respect to t into the parts (a, x_0), (x_0, ∞) where x_0 is chosen to satisfy (2.27) with λ set equal to zero therein, and write $J = J_1 + J_2$ where

$$J_1 = \int_0^c \frac{u\phi(x, u)\, du_1}{g(u)} \int_a^{x_0} f(t)\psi(t, u)\, dt, \tag{3.4}$$

$$J_2 = \int_0^c \frac{u\phi(x, u)\, du_1}{g(u)} \int_{x_0}^\infty f(t)\psi(t, u)\, dt. \tag{3.5}$$

First we consider J_2. On applying the Schwarz inequality to the t-integral in J_2, we find that

$$\int_{x_0}^\infty |f(t)\psi(t, u)|\, dt \leqq \varepsilon \left(\int_{x_0}^\infty |\psi|^2\, dx \right)^{1/2} \leqq \varepsilon \|\psi\|$$

where $\|\psi\|$ denotes the norm based on the interval (a, ∞) and is obtainable from (A12). On estimating the values of $\psi(a, u)$, $\psi_x(a, u)$ appearing in (A12) with the aid of (1.8) and (1.11) we find since $B \sim (n + \frac{1}{4})\pi$ that

$$\int_{x_0}^\infty |f(t)\psi(t, u)|\, dt \leqq \varepsilon \left(\frac{2}{\pi R u_1} \right)^{1/2} \left| \exp\left(\frac{1}{2} i(u - \nu)\pi \right) \right| \cdot [1 + \sinh 2A + O(u^{-1} e^A)]^{1/2}. \tag{3.6}$$

Similarly, we find that

$$|g(u)| \sim 2 \left| \frac{u}{\pi} \right|^{1/2} \left| \exp\left(\frac{1}{2} i(u - \nu)\pi \right) \right| \cdot [\cosh 2A + O(u^{-1} e^A)]^{1/2}. \tag{3.7}$$

It follows from (3.6) and (3.7) that

$$\left| \frac{1}{g(u)} \int_{x_0}^\infty f(t)\psi(t, u)\, dt \right| \leqq \varepsilon \left| \frac{1}{2} R u u_1 \right|^{-1/2} = \varepsilon \left(\frac{1}{2} R u_1 \right)^{-1/2} (R^2 + u_1^2)^{-1/4}.$$

On taking the modulus of (3.5) and using the above result and the fact that $\phi(x, u)$ is, by (2.4), bounded in the strip $0 \leqq \mathrm{Re}\,(u) \leqq c$, it follows that

$$|J_2| \leqq M\varepsilon \left(\frac{2}{R} \right)^{1/2} \int_0^c u_1^{-1/2} (R^2 + u_1^2)^{1/4}\, du_1 \tag{3.8}$$

where M is a constant. On taking u_1/R as new variable of integration in (3.8), it is seen that the integral appearing in the latter equation is $O(R^{1/2})$ as $R \to \infty$ so that $|J_2| \leqq M'\varepsilon$, where M' is a further constant, independent of R.

We now turn our attention to the quantity J_1 defined by (3.4) and which, by (2.15), may be written as

$$J_1 = \int_a^{x_0} f(t)\, dt \int_0^c \left[h(u, x, t) - \frac{1}{2} e^{u(x-t)} - \frac{1}{2} e^{u(2a-x-t)} \right] du_1. \tag{3.9}$$

Since $h = O(u^{-1} e^{u(x-t)})$ and this is simply $O(R^{-1})$ uniformly in t, since $u = u_1 + iR$ on the path of integration, the contribution of the function $h(u, x, t)$ to the value of the repeated integral on the right-hand side of (3.9) is $O(R^{-1})$. In regard to the first

exponential term on the right-hand side of (3.9), we find on carrying out the integration with respect to u_1, that the contribution of this term equals the integral

$$\int_a^{x_0} [e^{c(x-t)}-1](x-t)^{-1}\, e^{iR(x-t)} f(t)\, dt.$$

This integral tends to zero as $R\to\infty$ by the Riemann-Lebesgue lemma. A similar argument reveals that the contribution of the second exponential term present on the right-hand side of (3.9) vanishes in the limit as $R\to\infty$. Thus as $R\to\infty$, $J_1\to 0$ and $J\to J_2$, which is arbitrarily small so that $J\to 0$ as $R\to\infty$. Therefore the integral along the top side vanishes in the limit as this side recedes to infinity. A similar result holds for the integral along the bottom side of the rectangle, and this justifies the derivation of the expansion (3.1).

Appendix. In this Appendix the bounds on $\|e^{\lambda t}\psi(t,u)\|$ utilized in § 2 are derived. In addition some properties of the zeros of the function $g(u)$ are established. The required bounds may be obtained from the differential equation (1.1) by multiplying the latter by $e^{\lambda x}\bar{y}$, where $\bar{y}$ denotes the complex conjugate of y and $\lambda \geqq 0$, and integrating by parts, this leads to

$$\begin{aligned}&\int_a^\infty e^{\lambda x}(u^2+2h^2\cosh 2x)|y|^2\,dx\\ &\quad=-e^{\lambda a}\bar{y}(a)y_x(a)-\lambda\int_a^\infty e^{\lambda x}\bar{y}y_x\,dx-\int_a^\infty e^{\lambda x}|y_x|^2\,dx.\end{aligned}\tag{A1}$$

Further relations may be obtained by noting that $2\,\mathrm{Re}\,(y_x\bar{y}_{xx})=y_x\bar{y}_{xx}+\bar{y}_x y_{xx}$ so that

$$2\,\mathrm{Re}\int_a^\infty e^{\lambda x}y_x\bar{y}_{xx}\,dx=-e^{\lambda a}|y_x(a)|^2-\lambda\int_a^\infty e^{\lambda x}|y_x|^2\,dx.\tag{A2}$$

Similarly, we find that

$$2\,\mathrm{Re}\int_a^\infty e^{\lambda x}\bar{y}y_x\,dx=-e^{\lambda a}|y(a)|^2-\lambda\int_a^\infty e^{\lambda x}|y|^2\,dx,\tag{A3}$$

$$\begin{aligned}&2\,\mathrm{Re}\int_a^\infty e^{\lambda x}\cosh 2x\bar{y}y_x\,dx\\ &\quad=-e^{\lambda a}\cosh 2a|y(a)|^2-\int_a^\infty e^{\lambda x}(\lambda\cosh 2x+2\sinh 2x)|y|^2\,dx.\end{aligned}\tag{A4}$$

Setting $\bar{y}_{xx}=(u^2+2h^2\cosh 2x)\bar{y}$ in (A2), we obtain

$$2\,\mathrm{Re}\int_a^\infty e^{\lambda x}(u^2+2h^2\cosh 2x)\bar{y}y_x\,dx=-e^{\lambda a}|y_x(a)|^2-\lambda\int_a^\infty e^{\lambda x}|y_x|^2\,dx.\tag{A5}$$

Eliminating the integral involving the term $e^{\lambda x}\cosh 2x\bar{y}y_x$ between (A4) and (A5) yields

$$\begin{aligned}&2(u_1^2-u_2^2)\,\mathrm{Re}\int_a^\infty e^{\lambda x}\bar{y}y_x\,dx+4u_1u_2\,\mathrm{Im}\int_a^\infty e^{\lambda x}\bar{y}y_x\,dx\\ &\quad=-e^{\lambda a}|y_x(a)|^2+2h^2e^{\lambda a}\cosh 2a|y(a)|^2-\lambda\int_a^\infty e^{\lambda x}|y_x|^2\,dx\\ &\qquad+2h^2\int_a^\infty e^{\lambda x}(\lambda\cosh 2x+2\sinh 2x)|y|^2\,dx.\end{aligned}\tag{A6}$$

On extracting the imaginary parts, it follows from (A1) that

$$\lambda \operatorname{Im} \int_a^\infty e^{\lambda x} \bar{y} y_x \, dx = -e^{\lambda a} \operatorname{Im} \bar{y}(a) y_x(a) - 2u_1 u_2 \int_a^\infty e^{\lambda x} |y|^2 \, dx. \tag{A7}$$

On taking the real part of equation (A1) and then eliminating the integrals involving the terms $\operatorname{Re} \bar{y} y_x$, $\operatorname{Im} \bar{y} y_x$, and $|y_x|^2$ that occur in (A1), (A3), (A6), and (A7), after some reduction we find

$$\begin{aligned} &\left(u_1^2 - \frac{1}{4}\lambda^2\right)\left(u_2^2 + \frac{1}{4}\lambda^2\right) \int_a^\infty e^{\lambda x} |y|^2 \, dx + \lambda h^2 \int^\infty e^{\lambda x} (\lambda \cosh 2x + \sinh 2x) |y|^2 \, dx \\ &= \frac{1}{8} \lambda \, e^{\lambda a} \left[\left(u_2^2 - u_1^2 + \frac{1}{2}\lambda^2 - 2h^2 \cosh 2a \right) |y(a)|^2 + |y_x(a)|^2 - \lambda \operatorname{Re} y(a) \bar{y}_x(a) \right] \\ &\qquad + \frac{1}{2} u_1 u_2 \, e^{\lambda a} \operatorname{Im} y(a) \bar{y}_x(a). \end{aligned} \tag{A8}$$

The function y present in this equation is now taken to be the basic solution $\psi(x, u)$ introduced in § 1. To obtain a simple expression for the desired bound on the quantity $\|e^{\lambda x} \psi(x, u)\|$, first we consider large values of u. For such values we may employ (1.8) to show that

$$\operatorname{Re} \psi(x, u) \bar{\psi}_x(x, u) = -\frac{4}{\pi} |e^{i(u-v)\pi}| [\cos \theta \sinh 2A - \sin \theta \sin 2B + O(u^{-1} e^A)].$$

The expression on the right-hand side of this equation is negative for (large) values of u in the domain $\operatorname{Re}(u) \geqq c > 0$ since, by (2.12), A is large and positive there. Suppose now that we confine attention to the halfplane $u_1 \geqq \frac{1}{2}\lambda$; then, replacing λ by $2u_1$ in all of the terms appearing in the expression on the right-hand side of (A8) except the exponential ones, we show that the magnitude of this expression does not exceed that of the quantity

$$\begin{aligned} &\tfrac{1}{4} u_1 \, e^{\lambda a} [(u_2^2 + u_1^2) |\psi(a)|^2 + |\psi_x(a)|^2 - 2u_1 \operatorname{Re} \psi(a) \bar{\psi}_x(a) + 2u_2 \operatorname{Im} \psi(a) \bar{\psi}_x(a)] \\ &\quad = \tfrac{1}{4} u_1 \, e^{\lambda a} |u\psi(a) - \psi_x(a)|^2 \end{aligned}$$

where $\psi(a)$, $\psi_x(a)$ denote $\psi(a, u)$, $\psi_x(a, u)$, respectively. Equation (A8) now gives the inequality

$$\left(u_1^2 - \frac{1}{4}\lambda^2\right)\left(u_2^2 + \frac{1}{4}\lambda^2\right) \int_a^\infty e^{\lambda x} |\psi|^2 \, dx \leqq \frac{1}{4} u_1 \, e^{\lambda a} |u\psi(a) - \psi_x(a)|^2. \tag{A9}$$

This bound holds whenever u is large and $\operatorname{Re}(u) > \frac{1}{2}\lambda$. If u is not large, the bound (A9) still holds, by the method of derivation, at those parts of the halfplane $\operatorname{Re}(u) > \frac{1}{2}\lambda$ where the quantity $\operatorname{Re} \psi(a, u) \bar{\psi}_x(a, u)$ is negative. If this quantity is positive a slightly different bound applies, for then it is clear that the expression on the right-hand side of (A8) does not exceed the quantity

$$\begin{aligned} &\tfrac{1}{4} u_1 \, e^{\lambda a} [(u_2^2 + u_1^2) |\psi(a)|^2 + |\psi_x(a)|^2 + 2u_1 \operatorname{Re} \psi(a) \bar{\psi}_x(a) + 2u_2 \operatorname{Im} \psi(a) \bar{\psi}_x(a)] \\ &\quad = \tfrac{1}{4} u_1 \, e^{\lambda a} |\bar{u}\psi(a) + \psi_x(a)|^2 \end{aligned}$$

so that, in this case,

$$\left(u_1^2 - \frac{1}{4}\lambda^2\right)\left(u_2^2 + \frac{1}{4}\lambda^2\right) \int_a^\infty e^{\lambda x} |\psi|^2 \, dx \leqq \frac{1}{4} u_1 \, e^{\lambda a} |\bar{u}\psi(a) + \psi_x(a)|^2. \tag{A10}$$

It follows from (A9) and (A10) that

$$\text{(A11)} \qquad \left(u_1^2-\frac{1}{4}\lambda^2\right)\left(u_2^2+\frac{1}{4}\lambda^2\right)\int_a^\infty e^{\lambda x}|\psi|^2\,dx \leqq \frac{1}{4}u_1\,e^{\lambda a}(|u\psi(a)|+|\psi_x(a)|)^2.$$

This inequality holds throughout the halfplane $\operatorname{Re}(u)>\frac{1}{2}\lambda$, and on replacing λ by 2λ we obtain the result (2.9) used earlier.

The properties of the zeros of the function $g(u)$ may be deduced in the following manner, starting with (A7). This equation holds for all real values of λ, and on setting $\lambda=0$ it reduces, when y is replaced by ψ, to

$$\text{(A12)} \qquad 2u_1u_2\int_a^\infty |\psi|^2\,dx = -\operatorname{Im}\bar{\psi}(a,u)\psi_x(a,u).$$

If u is a zero of the function $g(u)=k\psi(a,u)+\psi_x(a,u)$, (A12) takes the form

$$2u_1u_2\int_a^\infty |\psi|^2\,dx = \operatorname{Im} k|\psi(a,u)|^2.$$

It follows at once from this equation that the product u_1u_2 has the same sign as $\operatorname{Im}(k)$ and cannot vanish unless k is real. Thus if $\operatorname{Im}(k)\neq 0$ there can be no real and no purely imaginary zeros. If $\operatorname{Im}(k)$ is positive, so is u_1u_2, and the zeros are confined to the first and third quadrants of the u-plane. However, if $\operatorname{Im}(k)$ is negative, the zeros must be positioned in the second and fourth quadrants.

We now consider the location of the large zeros and for this purpose suppose, as in the paper, that $\operatorname{Im}(k)>0$ in which case the zeros lie in the first and third quadrants of the complex u-plane. Those zeros of large magnitude that are located in the first quadrant may be obtained by substituting the asymptotic formulas (1.8) and (1.10) into $g(u)=k\psi(a,u)+\psi_x(a,u)$ and equating the result to zero. This leads to the equation

$$u\log\left(\frac{2u}{he}\right)-ua-\frac{1}{2}iu\pi-\frac{1}{4}i\pi = in\pi+O(u^{-1})$$

where n is a large enough positive integer. On setting $u=r_n e^{i\theta_n}$ and separating real and imaginary parts, we find

$$\text{(A13)} \qquad r_n\cos\theta_n\left[\log\left(\frac{2r_n}{he}\right)-a\right]+\left(\frac{\pi}{2}-\theta_n\right)r_n\sin\theta_n = O(r_n^{-1}),$$

$$\text{(A14)} \qquad r_n\sin\theta_n\left[\log\left(\frac{2r_n}{he}\right)-a\right]-\left(\frac{\pi}{2}-\theta_n\right)r_n\cos\theta_n = \left(n+\frac{1}{4}\right)\pi+O(r_n^{-1}).$$

Since there are no real and no purely imaginary zeros, the values $\theta_n=0,\ \pi/2$ are ruled out. Therefore both terms on the left-hand side of (A13) are positive and each must tend to zero as $r_n\to\infty$. Thus $r_n\cos\theta_n[\log(2r_n/he)-a]\to 0$ as $r_n\to\infty$, implying that $r_n\cos\theta_n\to 0$, that is, the zeros of successively larger and larger magnitudes approach the imaginary axis. It also follows that $\theta_n\to\pi/2$ as $n\to\infty$, and on using these results in (A13) and (A14) we obtain the approximate equations

$$\text{(A15)} \qquad \begin{aligned} r_n\left[\log\left(\frac{2r_n}{he}\right)-a\right] &\sim \left(n+\frac{1}{4}\right)\pi+O(r_n^{-1}),\\ \theta_n &\sim \frac{1}{2}\pi+O\left(\frac{r_n^{-2}}{\log r_n}\right).\end{aligned}$$

Acknowledgment. The author thanks the referee for corrections and comments that led to the extension of the validity of (2.5) to functions of exponential growth.

REFERENCES

[1] J. MEIXNER AND F. W. SCHAFKE, *Mathieuschefunktionen und Spharoidfunktionen*, Springer-Verlag, Berlin, 1954.

[2] D. NAYLOR, *On simplified asymptotic formulas for a class of Mathieu functions*, SIAM J. Math. Anal., 15 (1984), pp. 1205-1213.

[3] D. NAYLOR AND P. H. CHANG, *On a formula of inversion*, SIAM J. Math. Anal., 13 (1982), pp. 1053-1071.

[4] C. G. C. PITTS, *Asymptotic approximations to solutions of a second order differential equation*, Quart. J. Math. Oxford Ser. (2), 17 (1966), pp. 307-320.

[5] ———, *Simplified asymptotic approximations to solutions of a second order differential equation*, Quart. J. Math. Oxford Ser. (2), 21 (1970), pp. 223-242.

[6] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, Oxford University Press, London, 1950.

[7] ———, *Eigenfunction Expansions Associated with Second Order Differential Equations*, Vol. 1, Second edition, Oxford University Press, London, 1962.

[8] G. N. WATSON, *Theory of Bessel Functions*, Cambridge University Press, London, 1958.

SIAM J. MATH. ANAL.
Vol 20, No. 6, pp. 1514-1528, November 1989

016

TRANSFORMATIONS OF THE JACOBIAN AMPLITUDE FUNCTION AND ITS CALCULATION VIA THE ARITHMETIC-GEOMETRIC MEAN*

KENNETH L. SALA†

Abstract. With the aid of the Poisson summation formula, expressions for the Jacobian amplitude function, am $(z; m)$, along with the complete set of Jacobian elliptic functions are given that, aside from their branchpoints and poles, respectively, are convergent throughout the complex plane for arbitrary parameter m. By utilizing the expression for am $(z; m)$, its periodicity properties are determined in each of the regions $m<0$, $0<m<1$, and $m>1$. Novel yet fundamental identities are presented describing various linear and quadratic transformations of the Jacobian amplitude function. Finally, that method based on the arithmetic-geometric mean and most widely employed for calculating the Jacobian elliptic functions is shown to be, when interpreted explicitly in terms of am $(z; m)$ and its transformation properties, a method first and foremost for the calculation of the Jacobian amplitude and co-amplitude functions from which the elliptic functions themselves are subsequently evaluated by means of simple, trigonometric identities.

Key words. Jacobian amplitude function transformations, Jacobian elliptic functions, arithmetic-geometric mean

AMS(MOS) subject classifications. 33A25, 30D99, 41A58

1. Introduction.

The most familiar of the twelve-member family of Jacobian elliptic functions (JEF) is the copolar trio

$$sn(z; m) = \sin[\text{am}(z; m)] = \frac{\Theta_3}{\Theta_2}\frac{\Theta_1(z/\Theta_3^2; q)}{\Theta_4(z/\Theta_3^2; q)},$$

$$cn(z; m) = \cos[\text{am}(z; m)] = \frac{\Theta_4}{\Theta_2}\frac{\Theta_2(z/\Theta_3^2; q)}{\Theta_4(z/\Theta_3^2; q)}, \tag{1.1}$$

$$dn(z; m) = \frac{d}{dz}\text{am}(z; m) = \frac{\Theta_4}{\Theta_3}\frac{\Theta_3(z/\Theta_3^2; q)}{\Theta_4(z/\Theta_3^2; q)},$$

where am $(z; m)$ is the Jacobian amplitude function, m is the Jacobian parameter ($k=+m^{1/2}$ is the modulus), $q=\exp[-\pi K'(m)/K(m)]$ is the nome with $K(m)$ and $K'(m)=K(1-m)$ the Jacobian quarter periods, and $\Theta_i(z; q)$, $i=1, \cdots, 4$, are the theta functions with Θ_i denoting $\Theta_i(z=0; q)$. The remaining members of the JEF family can be defined directly either as reciprocals or ratios of these three functions or by adding to the argument z one or both of the quarter periods, e.g., $cd(z; m) = cn(z; m)/dn(z; m) = sn(z+K; m)$. In what follows we will assume that the parameter m is real but otherwise arbitrary while the variable $z=x+iy$ is, in general, arbitrary and complex. Comprehensive descriptions of elliptic functions and JEF in particular may be found in [8], [10], [20], and [23], while extensive compendia of the properties of JEF are given in [5], [13], [15], and [17]. In general, well-known identities involving JEF will be cited without specific reference since they may be found in any of the aforementioned works.

The canonical definitions of the JEF given by (1.1) represent two characteristically distinct approaches to the description of these functions. Historically, the JEF were first defined as inverses of elliptic integrals with the basis of this approach summarized

* Received by the editors September 15, 1986; accepted for publication (in revised form) September 12, 1988.

† Department of Communications, Communications Research Center, P.O. Box 11490, Station H, Ottawa, Ontario, Canada K2H 8S2.

by the fundamental identity $\text{am}(z; m) = F^{-1}(z; m)$ where $F(z; m)$ is the elliptic integral of the first kind (a historical account of the development of elliptic function theory is given by Alling [1]). However, the study of the JEF via the amplitude function is not and has never been the favored approach principally for the reason that, since $\text{am}(z; m)$ is not itself an elliptic function, this approach could not effectively exploit the many general and powerful theorems for elliptic functions but would instead be forced to rely almost exclusively on "brute force" algebraic methods. With origins traceable to Jacobi's seminal work *Fundamenta Nova* [12], the preferred approach to the study of the JEF has been through the theta functions, which, of course, are entire functions with simple zeros. In modern texts on elliptic function theory (e.g., Chandrasekharan [8]), the function $\text{am}(z; m)$ is ignored altogether.

To describe the amplitude function thus as one of the more obscure higher transcendental functions would be an understatement. The extent of its inconspicuousness is best illustrated with the example of the classical problem of the simple pendulum. The angular displacement Θ of a point mass μ constrained to swing in a vertical plane by a massless, rigid rod of length R is described by the nonlinear equation (Whittaker [22])

$$\frac{d^2\Theta}{dt^2} + \frac{g}{R}\sin\Theta = 0, \tag{1.2}$$

or, equivalently,

$$\frac{1}{2}\mu R^2\left[\frac{d\Theta}{dt}\right]^2 + \mu g R(1-\cos\Theta) = E_0 \tag{1.3}$$

where the total energy E_0 is a constant. Note that (1.2) is also identical in form to the traveling wave, sine-Gordon equation. With the most general possible initial conditions of $\Theta = 0$ and $d\Theta/dt = [2E_0/\mu R^2]^{1/2}$ for $t = 0$ (this choice places no restrictions on the value of E_0), the exact general solution to (1.2) and (1.3) is simply

$$\Theta(t) = 2 \cdot \text{am}\left[\left(\frac{g}{mR}\right)^{1/2} t; m\right], \qquad m = \frac{2\mu g R}{E_0} \tag{1.4}$$

a result that follows immediately from the identities $(d/dx)\text{am}(x; m) = dn(x; m)$ and $(d/dx)\, dn(x; m) = -(m/2)\sin[2\text{am}(x; m)]$. Despite the simplicity of this result, the explicit solution (1.4) has heretofore never been published even though dozens of texts and papers have treated the simple pendulum problem "exactly." Invariably, these "exact" treatments solve not explicitly for $\Theta(t)$ but rather for the variable $\sin(\Theta/2)$ (see, e.g., Whittaker [22] and Alling [1]) and, furthermore, choose to either ignore entirely the rotating ($m < 1$) pendulum by adopting initial conditions that restrict the value of m to $m > 1$, or to treat the cases of $m < 1$ and $m > 1$ as distinct problems (the special case of $m = 1$ is also often treated separately). The distinction, however, between $\Theta(t)$ and $\sin(\Theta/2)$ is not a trivial one; the latter is a true doubly periodic function for all values of the parameter $m \neq 1$ whereas the amplitude function possesses a real period if and only if $m > 1$, i.e., only the $\text{am}(z; m)$ solution as given in (1.4) explicitly and unequivocally distinguishes between the oscillating ($m > 1$) and rotating ($m < 1$) pendulum solutions. In addition, it is important to note that (1.4) is a solution to the pendulum equation (1.2) for arbitrary values of m, i.e., it is solely the initial conditions that determine the specific value of the parameter m. Thus we have, from (1.4), that $\sin(\Theta/2) = sn[(g/mR)^{1/2}t; m] = k^{-1}sn[(g/R)^{1/2}t; 1/m]$, revealing that both cases of a parameter greater than 1 and less than 1 (as well as $m = 1$) are succinctly and

completely delineated by the result in (1.4) and that, in contrast, the "traditional" division of this problem into two (or three) distinct cases is unnecessarily redundant.

In the following, utilizing a completely novel representation for am $(z; m)$ that is convergent throughout the complex plane excepting only the logarithmic branchpoints of the function (hereafter the term "unrestricted representation" will be used to denote any representation of a function that is valid throughout the complex plane except at any isolated, singular points and/or branchpoints of the function), we examine its periodicity for all real values of m. We also present various linear and quadratic transformations of the amplitude function corresponding to, e.g., the complementary parameter transformation, the Landen and Gauss transformations, etc. Although the expression of these transformations in terms of the JEF are well known, the results presented here for am $(z; m)$ are, with one exception, new results. As will be evident, the transformations for am $(z; m)$ offer concise, straightforward representations for these transformations and, in certain cases, offer a simple representation for which the corresponding JEF transformation is considerably more complicated. An example of the latter is the ascending Landen transformation that takes a simple form for am $(z; m)$ whereas the identities involving the JEF are algebraic. In addition, the formulae presented here offer further insight into the nature of these basic transformations beyond that associated strictly with the JEF formulae.

Principally for reasons of computational efficiency, the most widely used method for calculating the JEF (and elliptic integrals) is that based on the arithmetic-geometric mean along with various supplemental relations normally involving specific transformations directly related to the function to be evaluated (see the general articles by King [14], Carlson [6], and Milne-Thomson [17]). The term "arithmetic-geometric mean" will henceforth be understood to include whatever supplemental relations are used in conjunction with the arithmetic-geometric mean itself in the overall calculation of the specific function in question. The final section of this paper describes the method of the arithmetic-geometric mean explicitly in terms of the amplitude function and its transformation properties and will demonstrate that the method of the arithmetic-geometric mean is first and foremost a technique for the calculation **NOT** of the JEF but rather of am $(z; m)$ directly (along with the "coam" function am $(K-z; m)$). It is emphasized that the intent of this section is not to define or present algorithms for the arithmetic-geometric mean as applied to the calculation of the JEF; there exist several excellent, comprehensive descriptions of this technique [1], [6], [7], [9], [14], [16], including strictly algebraic versions [6], [21], computer algorithms [4], [11], as well as versions permitting complex parameters [9]. Rather, we wish to show that the actual basis for this technique is best and most clearly described in terms of the transformation formulae for am $(z; m)$ presented in the first parts of this paper.

2. Unrestricted representations for the Jacobian functions. The Fourier series for the functions $dn(z; m)$ and am $(z; m)$ may be written in the following form:

$$\begin{aligned} dn(z; m) &= \frac{\pi}{2K} + \frac{2\pi}{K} \sum_{n=1}^{\infty} \frac{q^n}{1+q^{2n}} \cos\left(\frac{n\pi z}{K}\right) \\ &= \frac{\pi}{2K} \sum_{n=-\infty}^{\infty} \operatorname{sech}\left[n\pi \frac{K'}{K}\right] e^{in\pi z/K} \end{aligned} \tag{2.1}$$

and

$$\operatorname{am}(z; m) = \int_0^z dn(z; m)\, dz = \frac{\pi z}{2K} + 2 \sum_{n=1}^{\infty} \frac{1}{n} \frac{q^n}{1+q^{2n}} \sin\left(\frac{n\pi z}{K}\right). \tag{2.2}$$

However, as a consequence of the fact that the variable z and the summation index n are cofactors in the Fourier series representations, these expressions are only valid throughout the restricted domain $|\mathrm{Im}\,(z/K)| < \mathrm{Im}\,(iK'/K)$. For example, for $0 < m < 1$ where both K and K' are real, these expressions are valid only in the infinite strip $|\mathrm{Im}\,(z)| < K'(m)$. In addition, or rather as a result of this limitation, the Fourier series such as that given by (2.1) for $dn(z; m)$ account explicitly only for the periodicity properties with respect to the quarter-period $K(m)$ and completely fail to describe the behavior with respect to the quarter-period $iK'(m)$. To arrive at an unrestricted representation for $dn(z; m)$, we apply the Poisson summation formula (see Bellman [2] for a discussion and examples of the applicability of this formula)

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(u)\, e^{2\pi i n u}\, du \right] \tag{2.3}$$

to the second of (2.1) with $f(n) = \mathrm{sech}\,[n\pi K'/K] \exp\{in\pi z/K\}$. Replacing the variable "n" with "u" and evaluating the integral given in (2.3) leads directly to the result

$$dn(z; m) = \frac{\pi}{2K'} \sum_{n=-\infty}^{\infty} \mathrm{sech}\left[\frac{\pi K}{K'} \left(n + \frac{z}{2K} \right) \right]. \tag{2.4}$$

This expression for the function $dn(z; m)$ is superior to the Fourier series representation (2.1) in that, (a) equation (2.4) is convergent throughout the entire complex plane, poles excepted, and, (b) partly as a consequence of this, it describes equally explicitly the periodicity of $dn(z; m)$ with respect to *both* $K(m)$ and $K'(m)$ (the actual periods are $2K(m)$ and $4iK'(m)$). Indeed, since the variable z and the summation index n appear as additive terms in (2.4) in contrast to the Fourier series, (2.1), where they are multiplicative factors, the *only* condition required to ensure convergence of the expression (2.4) is $\mathrm{Re}\,(K/K') \neq 0$ which, for real m, is equivalent to $m \neq 0$.

The analogous Poisson-sum-transformed expressions for $sn(z; m)$ and $cn(z; m)$, from which the remaining members of the JEF family are derived as noted previously, are found by following exactly similar procedures as for the case of $dn(z; m)$ above, i.e., the Fourier series for these functions are first converted to a summation over an index "n" running from $-\infty$ to $+\infty$, the summation term is substituted into (2.3), and, following a substitution of the variable "n," the integration is performed. The final results may be compactly expressed in the following form, with $A = \pi/2K'$:

$$dn(z; m) = A \sum_{n=-\infty}^{\infty} \mathrm{sech}\left[\frac{\pi K}{K'} \left(n + \frac{z}{2K} \right) \right],$$

$$k \cdot cn(z; m) = A \sum_{n=-\infty}^{\infty} (-1)^n \mathrm{sech}\left[\frac{\pi K}{K'} \left(n + \frac{z}{2K} \right) \right], \tag{2.5}$$

$$k \cdot sn(z; m) = A \sum_{n=-\infty}^{\infty} (-1)^n \tanh\left[\frac{\pi K}{K'} \left(n + \frac{z}{2K} \right) \right],$$

$$k' \cdot nd(z; m) = A \sum_{n=-\infty}^{\infty} \mathrm{sech}\left[\frac{\pi K}{K'} \left(n + \frac{1}{2} + \frac{z}{2K} \right) \right],$$

$$-kk' \cdot sd(z; m) = A \sum_{n=-\infty}^{\infty} (-1)^n \mathrm{sech}\left[\frac{\pi K}{K'} \left(n + \frac{1}{2} + \frac{z}{2K} \right) \right], \tag{2.6}$$

$$k \cdot cd(z; m) = A \sum_{n=-\infty}^{\infty} (-1)^n \tanh\left[\frac{\pi K}{K'} \left(n + \frac{1}{2} + \frac{z}{2K} \right) \right],$$

$$cs(z;\, m) = A \sum_{n=-\infty}^{\infty} \operatorname{csch}\left[\frac{\pi K}{K'}\left(n+\frac{z}{2K}\right)\right],$$

(2.7)
$$ds(z;\, m) = A \sum_{n=-\infty}^{\infty} (-1)^n \operatorname{csch}\left[\frac{\pi K}{K'}\left(n+\frac{z}{2K}\right)\right],$$

$$ns(z;\, m) = A \sum_{n=-\infty}^{\infty} (-1)^n \coth\left[\frac{\pi K}{K'}\left(n+\frac{z}{2K}\right)\right],$$

$$-k' \cdot sc(z;\, m) = A \sum_{n=-\infty}^{\infty} \operatorname{csch}\left[\frac{\pi K}{K'}\left(n+\frac{1}{2}+\frac{z}{2K}\right)\right],$$

(2.8)
$$k' \cdot nc(z;\, m) = A \sum_{n=-\infty}^{\infty} (-1)^n \operatorname{csch}\left[\frac{\pi K}{K'}\left(n+\frac{1}{2}+\frac{z}{2K}\right)\right],$$

$$dc(z;\, m) = A \sum_{n=-\infty}^{\infty} (-1)^n \coth\left[\frac{\pi K}{K'}\left(n+\frac{1}{2}+\frac{z}{2K}\right)\right].$$

The symmetry of these expressions is striking; with the exception of certain factors of $\pm i$, the right-hand sides of (2.5)-(2.8) are, interestingly, exactly the set of (symmetrical) primitive elliptic functions originally defined by Neville [18], [19]. All twelve of these expressions are valid throughout the complex plane for arbitrary $m \neq 0$, their respective poles excepted. Each of the numbered equations represents a copolar trio of the JEF while the three quartets formed from the respective members of each of these trios are coperiodic. The expressions for *sn*, *cn*, and *dn* recently have been presented and discussed by Boyd [3]. However, to the best of the author's knowledge, (2.5)-(2.8) for the complete JEF family have not been published previously.

Integration of (2.4) results in an expression for the amplitude function in the form

(2.9)
$$\operatorname{am}(z;\, m) = \sum_{n=-\infty}^{\infty} \operatorname{gd}\left[\frac{\pi K}{K'}\left(n+\frac{z}{2K}\right)\right]$$

(2.10)
$$= \operatorname{gd}\left[\frac{\pi z}{2K'}\right] + \sum_{n=1}^{\infty}\left[\operatorname{gd}\left[\frac{\pi K}{K'}\left(n+\frac{z}{2K}\right)\right] - \operatorname{gd}\left[\frac{\pi K}{K'}\left(n-\frac{z}{2K}\right)\right]\right]$$

where $\operatorname{gd}(z)$ is the Gudermannian function. Equations (2.9) and (2.10) converge throughout the complex plane except at the logarithmic branchpoints $z = 2sK + (2t+1)iK'$ where s and t are arbitrary integers. Equation (2.9) will serve as the basis for the derivation of the various identities in the following work so that the results obtained will be valid without restrictions on the range of z; those results obtained by direct reference to the analogous JEF relations, i.e., by "inversion," are generally accompanied by restrictions on the range of the real and/or imaginary parts of the variable z.

It is worthwhile noting that although the unrestricted representations given above for the JEF and for $\operatorname{am}(z;\, m)$ are much more attractive for analytical purposes than their limited Fourier series counterparts, neither set of expressions, for purposes of numerical calculation, is as computationally efficient as those methods based either on the arithmetic-geometric mean or on the use of the theta functions (cf. [4], [6], [11], [14]). High precision, numerical evaluation of the JEF or $\operatorname{am}(z;\, m)$ using the Fourier series is truly practical only when $m \approx 0$ or, using the expressions above, when $m \approx 1$.

Since the characteristics of the function $\operatorname{am}(z;\, m)$ in the complex plane are so intimately connected with the nature of the Gudermannian function, a brief accounting

of gd (z) is in order at this point. Because the function sech (z) is a singly periodic, meromorphic function with simple poles at the points $(2t+1)i\pi/2$ with residues of $(-1)^{t+1}i$ where $t=0, \pm1, \pm2, \cdots$, the Gudermannian function defined as the definite integral (Jahnke and Emde [13])

$$\text{gd}\,(z)=\alpha+i\beta=\int_0^z \text{sech}\,(z)\,dz, \qquad z\neq(2n+1)\frac{i\pi}{2},$$

where α and β are strictly real, is a single-valued, analytic function provided that the complex plane is cut along the branchlines lying on the imaginary axis from $(4t+1)i\pi/2$ to $(4t+3)i\pi/2$ (logarithmic branchpoints for gd (z)) where t is an arbitrary integer. Relations such as $\sinh(z)=\tan[\text{gd}\,(z)]$ and $\cosh(z)=\sec[\text{gd}\,(z)]$ follow directly from the definition above. The real and imaginary parts of gd (z) for $x\neq0$ are given explicitly and uniquely by the relations

$$\begin{aligned}\alpha&=\text{gd}\,(x)+\tan^{-1}[\text{csch}\,(x)]-\tan^{-1}[\cos(y)\,\text{csch}\,(x)],\\ \beta&=\tanh^{-1}[\sin(y)\,\text{sech}\,(x)]\end{aligned} \tag{2.11}$$

with $\text{gd}\,(x)=2\tan^{-1}[\tanh(x/2)]$ and where $|\alpha|<\pi$ and $|\alpha|\to\pi/2$, $|\beta|\to0$ as $|x|\to\infty$ for all y. For $x=0$ and $y\neq(2t+1)i\pi/2$, we have $\alpha=0$ and $\beta=\tanh^{-1}[\sin(y)]$. Note that gd (z) is singly periodic with period $2\pi i$ and that $\text{gd}\,(-z)=-\text{gd}\,(z)$ and $\text{gd}\,(z^*)=\text{gd}^*\,(z)$. Expanding sech (z) in terms of exp $(\pm z)$ and integrating leads to an unrestricted representation for the Gudermannian function in the form

$$\text{gd}\,(z)=\text{sgn}\,(x)\frac{\pi}{2}-2\sum_{n=0}^{\infty}\frac{(-1)^n}{2n+1}[\text{sgn}\,(x)\cosh[(2n+1)z]-\sinh[(2n+1)z]], \tag{2.12}$$

which is convergent throughout the complex plane, the logarithmic branchpoints $z=(2t+1)i\pi/2$ excepted, and where sgn $(x)=+1, 0$, or -1 according to $x>0$, $x=0$, or $x<0$, respectively (the real part of gd (z) vanishes along the imaginary axis). It follows directly from (2.12) that

$$\text{gd}\,(x+iy\pm in\pi)=\text{sgn}\,(x)\frac{1-(-1)^n}{2}\pi+(-1)^n\,\text{gd}\,(x+iy) \tag{2.13}$$

revealing a finite discontinuity (of 2π) in the real part of gd (z) across each of the branchcuts $(x=0, \cos(y)<0)$.

Before proceeding to examine the specific properties of the Jacobian amplitude function, it is appropriate at this point to discuss briefly its general characteristics in the complex plane given the basic results immediately above. Neville [18, pp. 18–20] has shown that the integral of an elliptic function having zero residues defines a doubly (additive) pseudoperiodic, meromorphic function. The function am $(z; m)$, in contrast, as the integral of the elliptic function $dn(z; m)$ having nonzero residues (specifically, $dn(z; m)$ has simple poles with residues of $-i$ at the points $2sK+(4t+1)iK'$ and $+i$ at the poles $2sK+(4t-1)iK'$, where s, t are arbitrary integers), is, in general, a doubly pseudoperiodic function with logarithmic branchpoints. Thus am $(z; m)$, were it defined solely by (1.1) and (2.2), would be an infinitely multiple-valued function of z with branches differing by integral multiples of 2π corresponding to the infinite number of possible paths of integration from zero to z encircling the poles of $dn(z; m)$ in different ways. However, by cutting the complex plane along the line segments joining these logarithmic branchpoints, specifically from $2sK+(4t+1)iK'$ to $2sK+(4t+3)iK'$, where s, t are integers, the function am $(z; m)$ is made single-valued and analytic throughout the cut, complex plane. Finally, a principal branch is selected from among

these single-valued branches by the requirement $\operatorname{am}(z=0; m)=0$. The real part of $\operatorname{am}(z; m)$ will be discontinuous (by 2π) across each of the branchlines. In nearly all respects, as the form of (2.9) intimates, the Jacobian amplitude function $\operatorname{am}(z; m)$ may be effectively considered as a doubly pseudoperiodic generalization of the singly periodic Gudermannian function $\operatorname{gd}(z)$, noting in particular the degeneracy $\operatorname{am}(z; m=1)=\operatorname{gd}(z)$.

3. The periodicity properties of am $(z; m)$ for real m. Exactly as with the JEF, the behavior of the function $\operatorname{am}(z; m)$ is characteristically different in the three distinct regions: $-\infty<m<0$, $0<m<1$, and $1<m<\infty$. From (2.9), which is valid for all parameter values, it follows that the amplitude function is always at least singly periodic when $m\neq 0$, with a period of $4iK'(m)$. Exploiting the fact that the amplitude function for real m, like the JEF, is strictly real whenever z is real, expressions for am are given in the regions $m<0$ and $m>1$ that reflect this characteristic. Unless specifically noted, the degenerate cases of $\operatorname{am}(z; m=0)=z$ and $\operatorname{am}(z; m=1)=\operatorname{gd}(z)$ are generally excluded from the relations below. Finally, it should be noted that, in all of the work to follow, the numerical value of the parameter m will generally be restricted to $0<m<1$ and parameters that are less than zero or greater than 1 will then be expressed explicitly in terms of m, e.g., a parameter greater than 1 will be represented by $1/m$ where $0<m<1$.

3.1. am $(z; m)$ for $0<m<1$. An arbitrary point in the complex plane may be represented as

$$z+2sK+2itK'=x+iy+2sK+2itK', \qquad |x/2K|<1 \quad \text{and} \quad |y/K'|<1 \tag{3.1}$$

where s and t are integers and where, for $0<m<1$, both K and K' are real. Only the lines $y=(2t+1)K'$, which include the logarithmic branchpoints of $\operatorname{am}(z; m)$, are excluded from the representation (3.1). When we use (2.12) and the expression for $\operatorname{am}(z; m)$ given by (2.10), it is a straightforward task to derive the result:

$$\operatorname{am}(z+2sK+2itK'; m)=s\pi+\operatorname{sgn}(x)\frac{1-(-1)^t}{2}\pi+(-1)^t\operatorname{am}(z; m). \tag{3.2}$$

Thus we have $\operatorname{am}(z+2sK+2itK'; m)=\operatorname{am}(z; m)$ if and only if $s=0$ and t is even and so, for $0<m\leqq 1$, the function $\operatorname{am}(z; m)$ is a singly periodic function with the strictly imaginary period $4iK'$. Equation (3.2) corresponds exactly with the relief figures for $\operatorname{am}(z; m)$ given by Jahnke and Emde [13]. Note that the branchlines so clearly illustrated in those figures are also explicitly accounted for by (3.2). For example, taking $s=2$, $t=1$, and $z=x>0$, we have $\operatorname{am}(4K+2iK'+x; m)=3\pi-\operatorname{am}(x; m)$, whereas $\operatorname{am}(4K+2iK'-x; m)=\pi+\operatorname{am}(x; m)$.

3.2. am $(z; m)$ for $m<0$. Denoting a negative parameter (imaginary modulus) by the expression $-m/m'$ where $m'=1-m$ and $0<m<1$, we have the identities [5], [10], [15], [23]

$$K(-m/m')=k'K(m), \qquad K'(-m/m')=k'K'(m)+ik'K(m). \tag{3.3}$$

Note that the sign used in this identity for $K'(-m/m')$ is ambiguous for real m; either a + or − may be used (consistently) without affecting the validity of the final results [19, pp. 103–107]. The expression (2.9) for the amplitude function with negative parameter then takes the following form:

$$\operatorname{am}\left(z; -\frac{m}{m'}\right)=\sum_{n=-\infty}^{\infty}\operatorname{gd}\left[\frac{\pi K}{K'+iK}\left(n+\frac{z}{2k'K}\right)\right]. \tag{3.4}$$

This expression is cumbersome in that the individual terms in the summation are complex when z is strictly real even though the function $\operatorname{am}(z; -m/m')$ itself is real in such a case. To overcome this shortcoming, consider the expression

$$G(z) = \sum_{n=-\infty}^{\infty} \left[(1+ia) \operatorname{sech}\left[a\pi\left(z+n+\frac{1}{2}\right)\right] - \operatorname{sech}\left[\frac{a\pi}{1+ia}(z+n)\right]\right]$$

where $a = K/K'$. This function $G(z)$ is a doubly periodic function since $G(z+1) = G(z+2i/a) = G(z)$ with possible poles at the isolated points $z_{st} = (s+\frac{1}{2}) + (2t+1)i/2a$, i.e., $G(z)$ is an elliptic function. It is, however, straightforward to prove that the limit $(z - z_{st}) \cdot G(z) = 0$ as $z \to z_{st}$ so that the function $G(z)$ is indeed without poles. As a consequence of Liouville's theorem, an elliptic function without poles must be a constant and, in fact, we have that $G(z) \equiv 0$. By integration we arrive at the desired result

$$\operatorname{am}\left(z; -\frac{m}{m'}\right) = \sum_{n=-\infty}^{\infty} \operatorname{gd}\left[\frac{\pi K}{K'}\left(n+\frac{1}{2}+\frac{z}{2k'K}\right)\right] - \frac{\pi}{2} \tag{3.5}$$

for which the only complex dependence is implicitly through z. Representing an arbitrary point in the complex plane as above with $|x/2k'K|$ and $|y/k'K'| < 1$ and using the representation (2.12) in (3.5) leads to the result

$$\begin{aligned} &\operatorname{am}\left[z \pm k'K + 2sK\left(-\frac{m}{m'}\right) + 2itK'\left(-\frac{m}{m'}\right); -\frac{m}{m'}\right] \\ &\quad = (s-t)\pi + (1 \pm \operatorname{sgn}(x))\frac{1-(-1)^t}{2}\pi \\ &\qquad + (-1)^t \operatorname{am}\left(z \pm k'K; -\frac{m}{m'}\right). \end{aligned} \tag{3.6}$$

The left-hand side of (3.6) will equal $\operatorname{am}(z \pm k'K; -m/m')$ if and only if $s = t = 2L$ so that $\operatorname{am}(z; -m/m')$ is singly periodic with the strictly imaginary period $4K(-m/m') + 4iK'(-m/m') = 4ik'K'$.

3.3. am $(z; m)$ for $m > 1$. Denoting a parameter greater than 1 by $1/m$ where $0 < m < 1$, we have [5], [10], [15], [23]

$$K(1/m) = kK(m) + ikK'(m), \qquad K'(1/m) = kK'(m) \tag{3.7}$$

noting that, exactly as for (3.3), the sign on the right-hand side of this equation is arbitrary (Neville [19, pp. 103–107]). The expression for $\operatorname{am}(z; 1/m)$ from (2.9) becomes

$$\begin{aligned} \operatorname{am}\left(z; \frac{1}{m}\right) &= \sum_{n=-\infty}^{\infty} \operatorname{gd}\left[\frac{\pi K}{K'}\left(n+\frac{z}{2kK}\right) + in\pi\right] \\ &= \sum_{n=-\infty}^{\infty} (-1)^n \operatorname{gd}\left[\frac{\pi K}{K'}\left(n+\frac{z}{2kK}\right)\right]. \end{aligned} \tag{3.8}$$

Then, with $|x/2kK|$ and $|y/kK'| < 1$, we find, using (2.13),

$$\begin{aligned} &\operatorname{am}\left(z + 2sK\left(\frac{1}{m}\right) + 2itK'\left(\frac{1}{m}\right); \frac{1}{m}\right) \\ &\quad = (-1)^s \operatorname{sgn}(x)\frac{1-(-1)^{s+t}}{2}\pi + (-1)^t \operatorname{am}\left(z; \frac{1}{m}\right). \end{aligned} \tag{3.9}$$

Thus we find that am $(z; 1/m)$ is a *doubly* periodic function, i.e., am $(z+2sK(1/m)+2itK'(1/m); 1/m) = \text{am}\,(z; 1/m)$ if *both* s and t are, independently, even integers. One of the periods, $4ikK'$, is strictly imaginary, while the other, $4kK$, is strictly real.

4. Linear and quadratic transformations of the amplitude function. This section will present the (linear) negative parameter, reciprocal parameter, and complementary parameter transformations as well as the (quadratic) Landen and Gauss transformations of the Jacobian amplitude function. Although only two linear transformations plus the Landen transformation are strictly necessary since the remaining linear and quadratic transformations can then be derived from these [8], [10], [20], all of the above-mentioned transformations are included here since they are the most familiar and widely used of the JEF transformations. The convention followed here for the nomenclature of the quadratic transformations is that defined by Carlson [6] for which the variable changes in the same/opposite manner as the parameter for the Gauss/Landen transformations. Concise yet general discussions of the transformation theory of elliptic functions may be found in the texts by Erdélyi et al. [10], Chandrasekharan [8], and Rauch and Lebowitz [20].

4.1. The negative parameter (imaginary modulus) transformation. The relationship between am $(z; m)$ and am $(z; -m/m')$ follows immediately from the identity given in (3.5), i.e.,

$$\text{am}\left(k'z; -\frac{m}{m'}\right) = \frac{\pi}{2} - \text{am}\,(K-z; m), \qquad -\infty < m < 1 \tag{4.1}$$

from which follow directly the JEF transformations $dn(k'z; -m/m') = nd(z; m)$, and so on. This transformation was first given by Jacobi in *Fundamenta Nova* [12, p. 90] and, apparently, subsequently forgotten. The amplitude function in the region $m<0$ is characteristically very similar to am $(z; m)$ with $0<m<1$. In each case, the function is singly periodic with an imaginary period $= 4i \cdot \text{Re}\,[K'(m)]$ while, for z strictly real, both am $(x; m)$ and am $(x; -m/m')$ are unbounded, monotonically increasing functions and so are invertible over the entire real axis. Finally, we note that limit am $(z; m) \equiv 0$ as $m \to -\infty$.

4.2. The complementary parameter (imaginary argument) transformation. Replacing m with m' in (2.9) gives immediately

$$\text{am}\,(z; m') = \sum_{n=-\infty}^{\infty} \text{gd}\left[\frac{\pi K'}{K}\left(n+\frac{z}{2K'}\right)\right]. \tag{4.2}$$

However, to relate am $(z; m)$ directly to am $(z; m')$, the familiar transformation for $dn(z; m')$ is rewritten as [5], [10], [23]

$$dn(z; m') = \frac{d}{dz}\,\text{am}\,(z; m') = \frac{dn(iz; m)}{cn(iz; m)} = \frac{-i}{\cosh\,(i\Phi)}\,\frac{d\Phi}{dz} = \frac{d}{dz}\,\text{gd}\,(-i\Phi) \tag{4.3}$$

where $\phi = \text{am}\,(iz; m)$. Integrating from zero to z yields the result

$$\text{am}\,(z; m') = \text{gd}\,[-i \cdot \text{am}\,(iz; m)], \qquad 0 \leqq m \leqq 1 \tag{4.4}$$

where $|\text{Re}\,(z)| < K'(m)$ and $|\text{Im}\,(z)| < K(m)$. Equations (3.1) and (3.2) may be used to extend the applicability of this result for arbitrary values of z. In certain respects, this transformation could be aptly subtitled the "circular-hyperbolic transformation" since it relates the "nearly circular" JEF to their "nearly hyperbolic" counterparts, i.e., the amplitude and elliptic functions with $m \approx 0$ to those with $m \approx 1$. In the extreme

limit of $m=0$, (4.4) gives $\operatorname{am}(z;1)=\operatorname{gd}(z)$. Finally, note that the parameters m and m' are interchangeable in (4.4) so that $\operatorname{am}(z;m)=\operatorname{gd}[-i\cdot\operatorname{am}(iz;m')]$.

4.3. The descending Landen transformation. Dealing for the moment with general, iterative subscripts, we define, for $0\leqq m_i\leqq 1$, the transformations $m_{i+1}\leftrightarrow m_i$ as

$$m_{i+1}=f_-(m_i)=\left[\frac{1-k_i'}{1+k_i'}\right]^2,\qquad m_i=f_+(m_{i+1})=\frac{4k_{i+1}}{(1+k_{i+1})^2} \tag{4.5}$$

such that $0\leqq m_{i+1}\leqq m_i\leqq 1$ and $f_-[f_+(m)]=f_+[f_-(m)]=m$. The quarter periods for the two parameters connected by f_- and f_+ are related as

$$K_{i+1}=\frac{1+k_i'}{2}K_i \quad\text{and}\quad K_{i+1}'=(1+k_i')K_i' \tag{4.6}$$

so that

$$\frac{K_{i+1}'}{K_{i+1}}=2\frac{K_i'}{K_i} \quad\text{and}\quad q_{i+1}=q_i^2. \tag{4.7}$$

Setting $i=0$ and using the notation $m_0=m$, we can write the identities

$$\operatorname{am}(z;m)=\sum_{n=-\infty}^{\infty}\operatorname{gd}\left[\frac{\pi K_1}{K_1'}\left(2n+(1+k')\frac{z}{2K_1}\right)\right] \tag{4.8}$$

and

$$\operatorname{am}(K-z;m)=\sum_{n=-\infty}^{\infty}\operatorname{gd}\left[\frac{\pi K_1}{K_1'}\left(2n+1-(1+k')\frac{z}{2K_1}\right)\right]. \tag{4.9}$$

Combining these equations gives the descending Landen transformation for the amplitude function

$$\operatorname{am}[(1+k')z;m_1]=\operatorname{am}(z;m)-\operatorname{am}(K-z;m)+\frac{\pi}{2},\qquad 0\leqq m<1. \tag{4.10}$$

4.4. The reciprocal parameter transformation. Using the relations for the quarter periods, (3.7), along with (4.8) and (4.9) above, we have directly

$$\operatorname{am}\left[(1-k')z;\frac{1}{m_1}\right]=\operatorname{am}(z;m)+\operatorname{am}(K-z;m)-\frac{\pi}{2},\qquad 0<m<1. \tag{4.11}$$

Up to this point, all of the transformations given for $\operatorname{am}(z;m)$ involve precisely those parameters that characterize the analogous JEF transformations. Although (4.11) is the correct, general form of the transformation relating the amplitude functions in the regions $0<m<1$ and $m>1$, it does not directly relate m to $1/m$ as its name suggests. To resolve this point, reference is made to (3.9) and the fact that, when and only when $m>1$, is $\operatorname{am}(z;m)$ a strictly oscillatory function with respect to *both* K and K'. Hence, except at its branchpoints, the function $\operatorname{am}(z;m>1)$ is bounded throughout the entire complex plane, and we may properly represent it as an inverse of some combination of JEF. From the familiar identity for $dn(u+v;m)$ it follows that, for $|\operatorname{Re}(u)/K|$, $|\operatorname{Re}(v)/K|<1$ and $|\operatorname{Im}(u)/K'|$, $|\operatorname{Im}(v)/K'|<1$,

$$\operatorname{am}(u+v;m)=\tan^{-1}[sc(u;m)\,dn(v;m)]+\tan^{-1}[dn(u;m)sc(v;m)]. \tag{4.12}$$

In particular, setting $u=kx$ and $v=iky$ leads to

$$\begin{aligned}\operatorname{am}(kx+iky;m)=&\tan^{-1}[sc(kx;m)dc(ky;m')]\\&+i\cdot\tanh^{-1}[dn(kx;m)sn(ky;m')].\end{aligned} \tag{4.13}$$

Transforming $m \to 1/m$ and rearranging terms within the brackets results in

$$\text{am}\left(kx+iky;\frac{1}{m}\right)=\sin^{-1}\left[\frac{k\cdot sn(x;\,m)}{[1-dn^2(x;\,m)sn^2(y;\,m')]^{1/2}}\right] + i\cdot\tanh^{-1}[k\cdot cn(x;\,m)sd(y;\,m')]. \tag{4.14}$$

In the strictest sense, relations such as these are incorrect whenever $1/m<1$ unless the values of the real and imaginary parts of the variable z are specifically restricted as stated above or as in (3.1) and (3.2). In contrast, however, the identity of (4.14) is valid *as written* when $1/m>1$ for arbitrary z, the branchpoints excepted. Thus, the reciprocal parameter transformation for the amplitude function, (4.14), can be rewritten succinctly (although essentially symbolically when z is complex) in the form

$$\text{am}\,(kz;\,1/m)=\sin^{-1}[k\cdot sn(z;\,m)],\qquad 0<m\leqq 1 \tag{4.15}$$

noting that limit am $(z;\,m)\equiv 0$ as $m\to\infty$. In particular, note that, for z real (or $y=4tK'$), the pragmatic identity am $(kx;\,1/m)=\sin^{-1}[k\cdot sn(x;\,m)]$ gives the correct value of the amplitude function for all values of x.

The similarity between the reciprocal parameter transformation in the form of (4.11) and the descending Landen transformation, (4.10), is remarkable and is, in part, related to the fact that $f_+(m_1)=f_+(1/m_1)$. By combining these two equations and invoking the result in (4.1), it is possible to write a completely general identity that relates the amplitude functions in the three regions of real m as follows:

$$\text{am}\,(z;\,m)=\text{am}\,(k'z;\,-m/m')+\text{am}\,((1-k')z;\,1/m_1) \tag{4.16}$$

with $(-m/m')\leqq 0\leqq m\leqq 1\leqq(1/m_1)$.

4.5. The ascending Landen transformation. Adding (4.10) and (4.11) leads immediately to the ascending Landen transformation for the amplitude function as follows:

$$\begin{aligned}\text{am}\,(z;\,m)&=\tfrac{1}{2}\,\text{am}\,[(1+k')z;\,m_1]+\tfrac{1}{2}\,\text{am}\,[(1-k')z;\,1/m_1]\\&=\tfrac{1}{2}\,\text{am}\,[(1+k')z;\,m_1]+\tfrac{1}{2}\sin^{-1}[k_1 sn[(1+k')z;\,m_1]].\end{aligned} \tag{4.17}$$

Note that many texts refer simply to "the Landen transformation," invariably meaning the descending Landen transformation corresponding to (4.10) above. The relationships for the JEF corresponding to (4.10) are rational ones [5], [23], whereas those corresponding to the ascending Landen transformation, (4.17), are algebraic relations. This is one instance in which, apart from its conciseness, the form of the amplitude function transformation is simple and straightforward in comparison to its JEF counterpart.

4.6. The ascending/descending Gauss transformations. The ascending/descending Gauss transformations are derived by combining the complementary parameter transformation with the descending/ascending Landen transformations [8], [10], [20]. From (4.4) and (4.10), for $0<m<1$, follows the ascending Gauss transformation in the form

$$\text{gd}\,[i\cdot\text{am}\,[(1+k_1)z;\,m]]=\text{gd}\,[i\cdot\text{am}\,(z;\,m_1)]+\text{gd}\,[i\cdot\text{am}\,(z+iK_1';\,m_1)]+\frac{\pi}{2} \tag{4.18}$$

while, from (4.4) and (4.17), the descending Gauss transformation is found to be

$$\begin{aligned}\text{gd}\,[i\cdot\text{am}\,(z;\,m_1)]&=\frac{1}{2}\,\text{gd}\,[i\cdot\text{am}\,(1+k_1)z;\,m]\\&\quad+\frac{1}{2}\,\text{gd}\left[i\frac{\pi}{2}-i\cdot\text{am}\,[K-(1+k_1)z;\,m]\right].\end{aligned} \tag{4.19}$$

As with the Landen transformations, many texts that refer simply to "the Gauss transformation" invariably mean the ascending Gauss transformation for which the transformation formulae for the JEF are rational expressions [5], [23], unlike those corresponding to the descending Gauss transformation for which the JEF identities are algebraic.

5. The method of the arithmetic-geometric mean and am $(x; m)$. Although, in principle, the method of the arithmetic-geometric mean (AGM) (or the theta functions) could be employed for a complex variable (and, conceivably, for complex m [9]), it is considerably more practical to calculate the real and imaginary parts of am $(z; m)$ and the JEF separately using identities such as (4.13). Thus, only strictly real variables x will hereinafter be considered. In addition, this section will deal with the "classical" method of the AGM (e.g., King [14]) utilizing various trigonometric or hyperbolic recursion identities as opposed to purely algebraic versions [6], [21]. The former, while nominally less efficient computationally, offer the advantage of calculating the true value of am $(x; m)$ for arbitrarily large $|x|$, i.e., including the contribution $s\pi$ given in (3.2). Moreover, while the relations between the method of the AGM and the transformations presented here hold true whichever version is adopted, these relations are more explicit and thus more readily recognized in the "classical" case.

The method of the AGM begins by iteratively calculating, with $0 < m < 1$, the trio of numbers

$$a_{n+1} = \tfrac{1}{2}(a_n + b_n), \quad b_{n+1} = (a_n b_n)^{1/2}, \quad c_{n+1} = \tfrac{1}{2}(a_n - b_n) \tag{5.1}$$

with starting values of $a_0 = 1$, $b_0 = k'$, and $c_0 = k$. The numbers a_n and b_n converge rapidly to a common limit $(= \pi/2K)$ while c_n, as a measure of the "error" with $c_n^2 = a_n^2 - b_n^2$, vanishes quadratically, i.e., $c_{n+1} = c_n^2/4a_{n+1}$. The calculation is stopped at the Nth step where, to some prescribed degree of accuracy, c_N is negligible. For the descending Landen version of the AGM, a sequence of angles $\phi_{N-1}, \phi_{N-2}, \cdots, \phi_0$ is then calculated sequentially using the recurrence relation

$$\sin(2\Phi_{n-1} - \Phi_n) = \frac{c_n}{a_n} \sin(\Phi_n) \quad \text{with } \Phi_N = 2^N a_N x \tag{5.2}$$

to which the amplitude function and the JEF are related as

$$\begin{aligned} &\text{am}(x; m) = \Phi_0, \qquad \text{am}(K - x; m) = \Phi_0 - \Phi_1 + \frac{\pi}{2}, \\ &sn(x; m) = \sin(\Phi_0), \qquad cn(x; m) = \cos(\Phi_0), \\ &dn(x; m) = \frac{cn(x; m)}{sn(K - x; m)} = \frac{\cos(\Phi_0)}{\cos(\Phi_0 - \Phi_1)}. \end{aligned} \tag{5.3}$$

Attention is drawn to the first two identities of (5.3) that, to the best of the author's knowledge, have not heretofore been given in any of the papers dealing with the AGM as a method of evaluating the JEF. Yet, given these two results, it follows that the evaluation of the JEF is, in fact, entirely incidental to this method, i.e., it is am $(x; m)$ and am $(K - x; m)$, which are the primary quantities found via the AGM from which the JEF are then calculated from simple trigonometric expressions.

To indeed establish the validity of the identities listed in (5.3), it is first noted that the (a_n, b_n, c_n) scale described by (5.1) is exactly equivalent to sequential applications

of the descending operator $f_-(m_i)$ of (4.5), i.e.,

$$
\begin{aligned}
&k_0 = k \\
&k_1 = (1-k_0')/(1+k_0') = f_-(m) = c_1/a_1 \\
&k_2 = (1-k_1')/(1+k_1') = f_-^2(m) = c_2/a_2 \\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
&k_{n+1} = (1-k_n')/(1+k_n') = f_-^{n+1}(m) = c_{n+1}/a_{n+1}
\end{aligned} \tag{5.4}
$$

and where $K(m) = (\pi/2)(1+k_1)(1+k_2)\cdots = (\pi/2a_N)$. With the identification of $k_n = c_n/a_n$, the recurrence relation (5.2) is immediately recognized as the ascending Landen transformation for the amplitude function, (4.17), so that the ϕ_n sequence is equivalent to

$$
\begin{aligned}
&\Phi_0 = \operatorname{am}(x; m) \\
&\Phi_1 = \operatorname{am}\left(\frac{2x}{(1+k_1)}; m_1\right) \\
&\Phi_2 = \operatorname{am}\left(\frac{4x}{(1+k_1)(1+k_2)}; m_2\right) \\
&\quad\vdots \qquad\qquad \vdots \\
&\Phi_n = \operatorname{am}\left(\frac{2^n x}{(1+k_1)(1+k_2)\cdots(1+k_n)}; m_n\right).
\end{aligned} \tag{5.5}
$$

Note that, even though it is the *descending* Landen transformation that is the basis of this particular version of the AGM and transforms the variable $\phi_0 \to \phi_N$ as immediately above, it is the *ascending* Landen transformation for $\operatorname{am}(x; m)$ that is used in the actual calculations to transform $\phi_N \to \phi_0$. With the aid of (4.10), the sequence of amplitude functions, (5.5), may be re-expressed as

$$
\begin{aligned}
\Phi_n = {} & \operatorname{am}(x; m) + (1-\delta_{n0})\left[\frac{\pi}{2} - \operatorname{am}(K-x; m)\right] \\
& + (1-\delta_{n0})(1-\delta_{n1}) \sum_{i=1}^{n-1} \sum_{j=1}^{2^{i-1}} \\
& \cdot \left[\operatorname{am}\left[\frac{2j-1}{2^i} K + x; m\right] - \operatorname{am}\left[\frac{2j-1}{2^i} K - x; m\right]\right]
\end{aligned} \tag{5.6}
$$

and

$$
\begin{aligned}
\Phi_n &= \frac{\pi}{2K} 2^n x + \sum_{j=1}^{\infty} \frac{2q^{2^n j}}{j(1+q^{2^{n+1}j})} \sin\left(\frac{2^n j\pi x}{K}\right) \\
&= \frac{\pi}{2K} 2^n x + \sum_{j=1}^{\infty} \frac{1}{j} \operatorname{sech}\left[\frac{2^n j\pi K'}{K}\right] \sin\left(\frac{2^n j\pi x}{K}\right).
\end{aligned} \tag{5.7}
$$

Equation (5.6), in particular, establishes the identities given in (5.3), while (5.7) offers explicit testimony to the extraordinarily rapid convergence inherent to the AGM. This latter equation states that, for the nth step, the first-order deviation of the variable ϕ_n from linearity will go as q raised to the 2^n power, i.e., given the relation of (4.7) for the nome, the "error" is reduced quadratically on each iteration.

The particular version of the AGM as just described with $b_0 = k'$ and the use of the recurrence relation (5.2) is referred to as the descending Landen version of the AGM. When we use $b_0 = k$ and (4.10) as the recurrence relation, the ascending Landen version of the AGM sequentially raises the parameter to unity where $\phi_N \approx \mathrm{gd}$. Ascending and descending versions of the AGM based on Gauss transformations are also possible (Carlson [6]). The use of an ascending and a descending transformation, in particular, allows the numerical range of m to be restricted to $0 < m \leqq \frac{1}{2}$. Whatever the particular version adopted, however, the calculation of the JEF from the final results, as for (5.3), proves to be incidental to the method in that the primary quantities that are calculated via the ϕ_n are the amplitude function $\mathrm{am}\,(x; m)$ along with the "coam" function $\mathrm{am}\,(K - x; m)$.

Calculation of $\mathrm{am}\,(x; m)$ along with the JEF for parameter values outside the range of $0 < m < 1$ may be done as directly and efficiently as for the case of $0 < m < 1$ through the use of the transformation formulae (4.1) and (4.11). To calculate the amplitude function and the JEF for the case of a negative parameter $-M$ where $0 < M < \infty$, the AGM is calculated as above with a parameter $m = M/(1+M)$ and a variable x/k' to give the results

$$\begin{gathered} \mathrm{am}\,(x; -M) = \Phi_1 - \Phi_0, \qquad \mathrm{am}\,(K(-M) - x; -M) = (\pi/2) - \Phi_0, \\ sn(x; -M) = \sin(\Phi_1 - \Phi_0), \qquad cn(x; -M) = \cos(\Phi_1 - \Phi_0), \\ dn(x; -M) = \cos(\Phi_1 - \Phi_0)/\cos(\Phi_0). \end{gathered} \tag{5.8}$$

To calculate the elliptic functions for the case of a parameter $M > 1$, the AGM as described above is calculated using a parameter $m = f_+(m_1)$, where $m_1 = 1/M$, and a variable $x/(1-k')$ to give the results

$$\begin{gathered} \mathrm{am}\,(x; M) = 2\Phi_0 - \Phi_1, \qquad \mathrm{am}\,(x/k_1; 1/M) = \Phi_1, \\ sn(x; M) = \sin(2\Phi_0 - \Phi_1), \qquad cn(x; M) = \cos(2\Phi_0 - \Phi_1), \\ dn(x; M) = \cos(\Phi_1). \end{gathered} \tag{5.9}$$

Note that, in the case of $M > 1$, it is not the coam function that is calculated along with $\mathrm{am}\,(x; M)$ but rather $\mathrm{am}\,(x/k_1; 1/M)$, which yields the value of $dn(x; M)$ directly. The directness of these algorithms, i.e., the calculation of the actual values of the amplitude function(s), may be contrasted with algorithms that calculate the JEF for parameters less than zero or greater than 1 as either rational or algebraic expressions involving values of JEF having $0 < m < 1$ (e.g., [4]).

6. Concluding remarks. The results presented in this paper have shown that the various linear and quadratic transformations of the JEF can be represented concisely by the corresponding transformation of the Jacobian amplitude function $\mathrm{am}\,(z; m)$. The nature of the amplitude function for arbitrary, real m has been shown to be markedly different according to whether $m < 1$ or $m > 1$, being a singly periodic function with a strictly imaginary period when $m < 1$ and $m \neq 0$ while, for $m > 1$, $\mathrm{am}\,(z; m)$ is a doubly periodic function with both a real and an imaginary period. Finally, the method of the arithmetic-geometric mean has been shown to be principally a method for the calculation of the functions $\mathrm{am}\,(x; m)$ and $\mathrm{am}\,(K - x; m)$ directly from which then follow the values of the various JEF.

Acknowledgments. The author thanks Dr. R. Somorjai of the National Research Council of Canada and Dr. M. Sablatash of the Communications Research Center, for many helpful discussions during the course of this work. I also extend my appreciation to the various individuals who refereed this paper, in particular the first referee, for their constructive criticisms and suggestions.

REFERENCES

[1] N. L. ALLING, *Real Elliptic Curves*, North-Holland, Amsterdam, 1981.

[2] R. BELLMAN, *A Brief Introduction to Theta Functions*, Holt, Rinehart, and Winston, New York, 1961.

[3] J. P. BOYD, *Cnoidal waves as exact sums of repeated solitary waves: new series for elliptic functions*, SIAM J. Appl. Math., 44 (1984), pp. 952-955.

[4] R. BULIRSCH, *Numerical calculation of elliptic integrals and elliptic functions*, Numer. Math., 7 (1965), pp. 78-90.

[5] P. F. BYRD AND M. D. FRIEDMAN, *Handbook of Elliptic Integrals for Engineers and Scientists*, Second edition, Springer-Verlag, Berlin, 1971.

[6] B. C. Carlson, *On computing elliptic integrals and functions*, J. Math. Phys., 44 (1965), pp. 36-51.

[7] ———, *Algorithms involving arithmetic and geometric means*, Amer. Math. Monthly, 78 (1971), pp. 496-505.

[8] K. CHANDRASEKHARAN, *Elliptic Functions*, Grundlehren Math. Wiss. 281, Springer-Verlag, Berlin, 1985.

[9] D. A. COX, *The arithmetic-geometric mean of Gauss*, L'Enseignement Math., 30 (1984), pp. 275-330.

[10] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, EDS., *Higher Transcendental Functions*, Vol. 2, Chap. 13, McGraw-Hill, New York, 1953.

[11] D. J. HOFSOMMER AND R. P. VAN DE RIET, *On the numerical calculation of elliptic integrals of the first and second kind and the elliptic functions of Jacobi*, Numer. Math., 5 (1963), pp. 291-302.

[12] C. G. JACOBI, *Fundamenta Nova*, Ponthieu, Paris, 1829.

[13] E. JAHNKE AND F. EMDE, *Tables of Functions with Formulas and Curves*, Dover, New York, 1945, pp. 41-106.

[14] L. V. KING, *On the Direct Numerical Calculation of Elliptic Functions and Integrals*, Cambridge University Press, Cambridge, UK, 1924.

[15] W. MAGNUS, F. OBERHETTINGER, AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Third edition, Springer-Verlag, Berlin, 1966.

[16] G. MIEL, *Of calculations past and present: the Archimedean algorithm*, Amer. Math. Monthly, 90 (1983), pp. 17-35.

[17] L. M. MILNE-THOMSON, *Jacobian elliptic functions and theta functions and elliptic integrals*, Chaps. 16 and 17 in Handbook of Mathematical Functions, M. Abramowitz and I. A. Stegun, eds., National Bureau of Standards, Washington, DC, 1964 (also Dover, New York, 1965).

[18] E. H. NEVILLE, *Jacobian Elliptic Functions*, Oxford University Press, London, 1944.

[19] ———, *Elliptic Functions: A Primer*, W. J. Langford, ed., Pergamon Press, Oxford, 1971.

[20] H. E. RAUCH AND A. LEBOWITZ, *Elliptic Functions, Theta Functions, and Riemann Surfaces*, Williams and Wilkins, Baltimore, MD, 1973.

[21] H. E. SALZER, *Quick calculation of Jacobian elliptic functions*, Comm. ACM, 5 (1962), p. 399.

[22] E. T. WHITTAKER, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Fourth edition, Dover, New York, 1944.

[23] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Fourth edition, MacMillan, New York, 1943.